Brisk guide to Mathematics

Jan Slovák

and

Martin Panák, Michal Bulant, Vladimir Ejov, Ray Booth

Brno, Adelaide, 2020

Authors:
Ray Booth
Michal Bulant
Vladimir Ezhov
Martin Panák
Jan Slovák

With further help of:
Aleš Návrat
Michal Veselý
...

Graphics and illustrations:
Petra Rychlá

Contents – practice

Contents – theory

Preface

The motivation for this textbook came from many years of lecturing Mathematics at the Faculty of Informatics at the Masaryk University in Brno. The programme requires introduction to genuine mathematical thinking and precision. The 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.

All these requirements have lead us to 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. This way, we want to encourage and help the readers to find their own way. Either to go through the examples and algorithms first, and then to come to explanations 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 percieve 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.

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, we are using icons to indicate complexity, difficulty etc.:

Further icons indicate unpleasant technicality and need of patiance, or possible entertainment and pleasure:
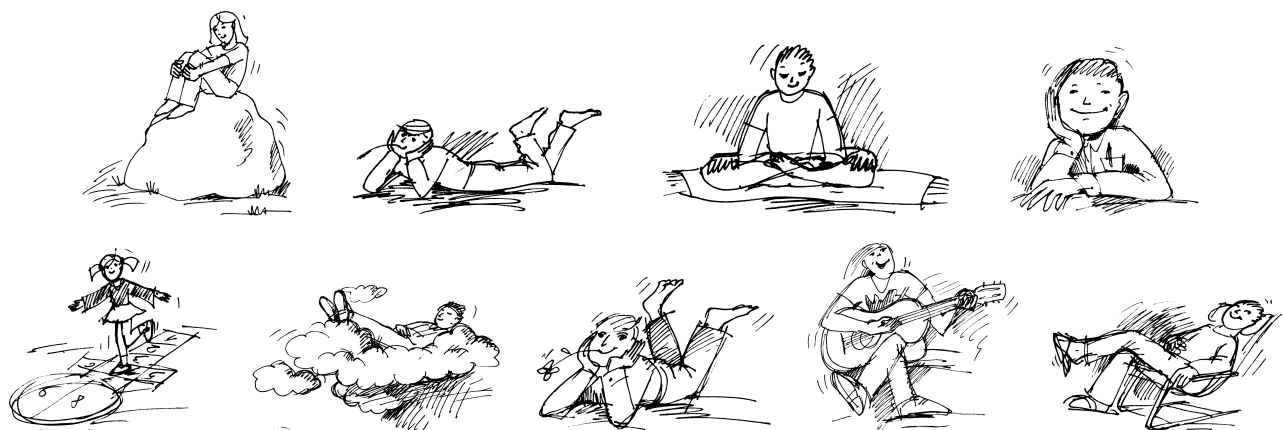
Similarly, we use various icons in the practical column:

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.B.4 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 percieved 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).
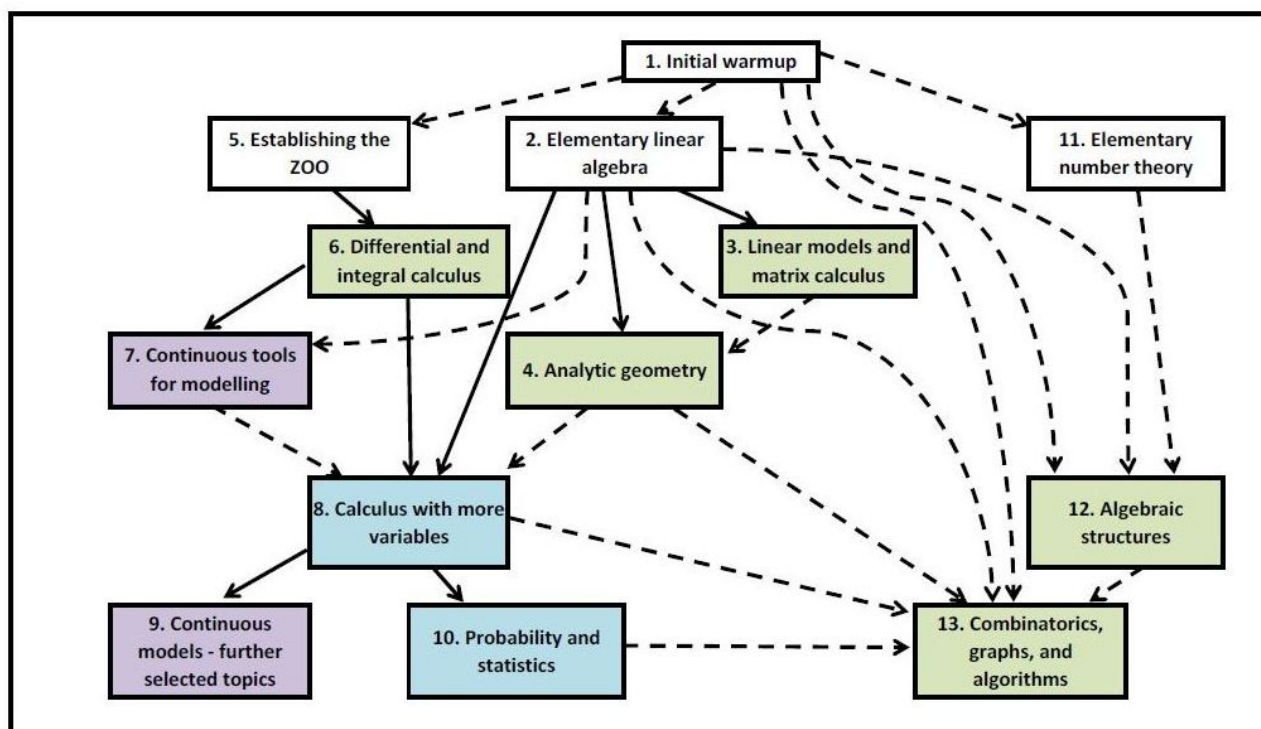
All topics covered in the book are now included (with more or less details) in our teaching of large four semester courses within our Mathematics minor programme, complemented by numerical seminars. In our teaching, the first semester covers chapters 1 and 2 and selected topics from chapters 3 and 4. 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 provids large parts of the chapters 11 through 13. Actually, the second semester could be offered in parallel with the first one, while the fourth semester could follow immediately after the first one. Probability and statistics (chapter 10) are offered as separate course in parallel.

Initial warmup

"value, difference, position"
– what it is and how to comprehend it?
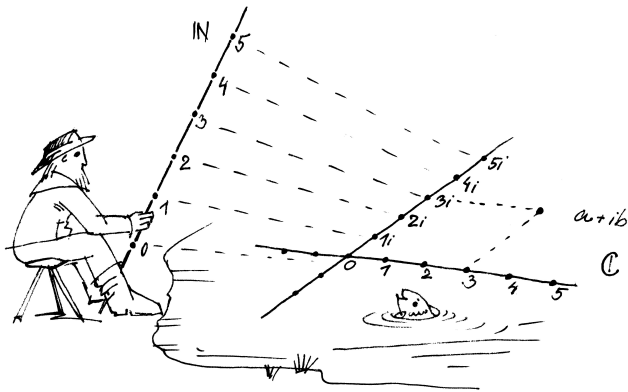
A. Numbers and functions

We can already work with natural, integer, rational and real numbers. We explain why rational numbers are not sufficient for us (although computers are actually not able to work with any other) and we recall the complex numbers (because even the real numbers are not adequate for some calculations).

1.A.1. Show that the integer 2 does not have a rational square root.

Solution. 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?

Assume we know that $(p/q)^2 = 2$ for natural numbers p and q that do not have common divisors greater than 1 (otherwise we can further reduce the fraction p/q). Then $p^2 = 2q^2$ is an even number. Thus, on the left-hand side p^2 is even. Therefore so is p because the alternative that p is odd would imply the contradiction that p^2 is odd. Hence, p is even and so p^2 is divisible by 4. So q^2 is even and so q must be even too. This implies that p and q both have 2 as a common factor, which is a contradiction. □

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. A full knowledge of mathematical thinking can be reached only through a long and complicated course of study.

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).

Perhaps sometimes our definitions and ideas will often 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 5 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

1.A.2. Remark. It can be proved that for all positive natural numbers n and x the n-th root $\sqrt[n]{x}$ of x is either natural or it is not rational, see 1.G.1.

All diagonals of squares irrational
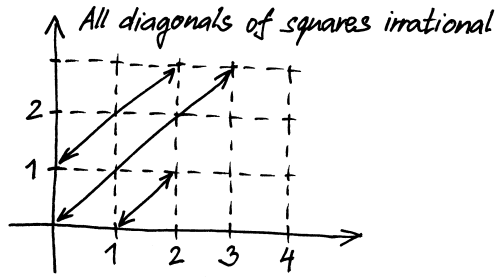
Next, we work out some examples with complex numbers. If you are not familiar with the basic concepts and properties of complex numbers, consult the paragraphs 1.1.3 through 1.1.4 in the other column.

1.A.3. Calculate $z_1 + z_2$, $z_1 \cdot z_2$, $\bar{z}_1$, $|z_2|$, $\frac{z_1}{z_2}$, for

a) $z_1 = 1 - 2i$, $z_2 = 4i - 3$;

b) $z_1 = 2$, $z_2 = i$.

Solution.

a) $z_1 + z_2 = 1 - 3 - 2i + 4i = -2 + 2i$, $z_1 \cdot z_2 = 1 \cdot (-3) - 8i^2 + 6i + 4i = 5 + 10i$, $\bar{z}_1 = 1 + 2i$, $|z_2| = \sqrt{4^2 + (-3)^2} = 5$, $\frac{z_1}{z_2} = \frac{z_1 \bar{z}_2}{|z_2|^2} = \frac{1 \cdot (-3) + 8i^2 + 6i - 4i}{25} = -\frac{11}{25} + \frac{2}{25}i$.

b) $z_1 + z_2 = 2 + i$, $z_1 \cdot z_2 = 2i$, $\bar{z}_1 = 2$, $|z_2| = 1$, $\frac{z_1}{z_2} = \frac{2}{i} = -2i$. $\square$

1.A.4. Determine

$$\left| \frac{(2+3i)\left(1 + i\sqrt{3}\right)}{1 - i\sqrt{3}} \right|.$$
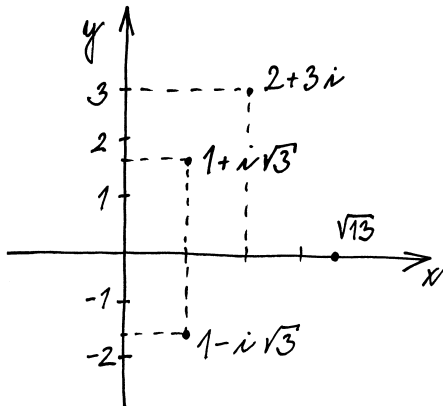
either considered to be a natural number, as is usual in computer science, or not a natural number as is usual in some other contexts. Thus the set of natural numbers is either $\mathbb{Z}^+$, or the set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. To count "one, two, three, ..." is learned already by children 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) $\qquad (a + b) + c = a + (b + c)$, for all a, b, c

(CG2) $\qquad a + b = b + a$, for all a, b

(CG3) $\qquad$ there exists 0 such that for all a, $a + 0 = a$

(CG4) $\qquad$ 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) $\qquad (a \cdot b) \cdot c = a \cdot (b \cdot c)$, for all a, b, c

(R2) $\qquad a \cdot b = b \cdot a$, for all a, b

(R3) $\qquad$ there exists 1 such that for all a $1 \cdot a = a$

(R4) $\qquad a \cdot (b + c) = a \cdot b + a \cdot c$, 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) $\qquad$ for every $a \neq 0$ there exists b such that $a \cdot b = 1$.

(ID) $\qquad$ 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 $\frac{1}{a}$.

Solution. Since the absolute value of the product (ratio) of any two complex numbers is the product (ratio) of their absolute values and every complex number has the same absolute value as its complex conjugate, we have that

$$\left|\tfrac{(2+3i)(1+i\sqrt{3})}{1-i\sqrt{3}}\right| = |2+3i| \cdot \tfrac{|1+i\sqrt{3}|}{|1-i\sqrt{3}|} = |2+3i|$$
$$= \sqrt{2^2+3^2} = \sqrt{13}. \qquad \square$$

1.A.5. Simplify the expression $\left(5\sqrt{3}+5i\right)^n$ for $n=2$ and $n=12$.

Solution. Using binomial theorem for $n=2$ we get

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

Taking powers one by one or doing an expansion using binomial theorem are in the case $n=12$ too much time-consuming. Let us rather write the number in polar form

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

and using de Moivre theorem we easily obtain

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

1.A.6. Determine the distance d of the numbers $z, \bar{z}$ in the complex plane for

$$z = \tfrac{\sqrt{3\sqrt{3}}}{2} - i\tfrac{3}{2}.$$

Solution. It is not difficult to realize that complex conjugates are in the complex plane symmetric with respect to the x-axis and the distance of a complex number from the x-axis equals its imaginary part. That gives $d=3$. $\qquad\square$

1.A.7. Express the number $z_1 = 2 + 3i$ in polar form. Express the number $z_2 = 3(\cos(\pi/3)+i\sin(\pi/3))$ in algebraic form.

Solution. The absolute value of $|z_1|$ (the distance of the point with Cartesian coordinates $[2,3]$ in the plane from the origin) is $\sqrt{2^2+3^2}=\sqrt{13}$. From the right triangle in the diagram we compute $\sin(\varphi)=3/\sqrt{13}$, $\cos(\varphi)=2/\sqrt{13}$. Thus $\varphi = \arcsin(3/\sqrt{13}) = \arccos(2/\sqrt{13}) \doteq 56.3°$. In total,

$$z_1 = \sqrt{13}\left(\tfrac{2}{\sqrt{13}} + i\cdot\tfrac{3}{\sqrt{13}}\right)$$
$$= \sqrt{13}\left(\cos\left(\arccos\left(\tfrac{2}{\sqrt{13}}\right)\right) + i\sin\left(\arcsin\left(\tfrac{3}{\sqrt{13}}\right)\right)\right).$$

Transition from polar form to algebraic form is even simpler:

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

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 $ab = 0$.

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 its 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.A.8. Express $z = \cos 0 + \cos \frac{\pi}{3} + i \sin \frac{\pi}{3}$ in polar form.

Solution. To express number z in polar form, we need to find its absolute value and argument. First we calculate the absolute value:

$$|z| = \sqrt{\left(\cos 0 + \cos \frac{\pi}{3}\right)^2 + \sin^2 \frac{\pi}{3}}$$
$$= \sqrt{\left(1 + \frac{1}{2}\right)^2 + \left(\frac{\sqrt{3}}{2}\right)^2} = \sqrt{3}.$$

For the argument φ, we have:

$$\cos \varphi = \frac{\operatorname{Re}(z)}{|z|} = \frac{1 + \frac{1}{2}}{\sqrt{3}} = \frac{\sqrt{3}}{2}, \quad \sin \varphi = \frac{\operatorname{Im}(z)}{|z|} = \frac{1}{2},$$

therefore $\varphi = \pi/6$. Thus

$$z = \sqrt{3}\left(\cos \frac{\pi}{6} + i \sin \frac{\pi}{6}\right). \qquad \square$$
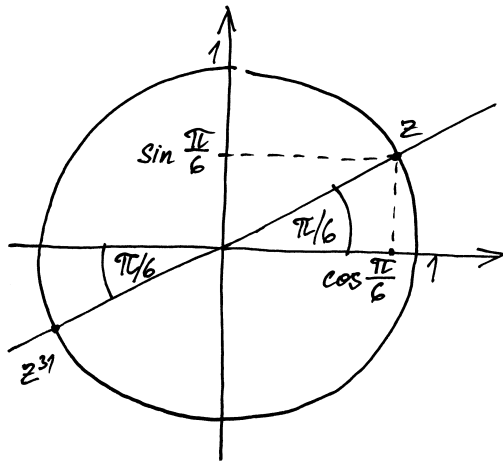
1.A.9. Using de Moivre theorem, calculate

$$\left(\cos \frac{\pi}{6} + i \sin \frac{\pi}{6}\right)^{31}.$$

Solution. We obtain

$$\left(\cos \frac{\pi}{6} + i \sin \frac{\pi}{6}\right)^{31}$$
$$= \cos \frac{31\pi}{6} + i \sin \frac{31\pi}{6} = \cos \frac{7\pi}{6} + i \sin \frac{7\pi}{6} = -\frac{\sqrt{3}}{2} - i \frac{1}{2}.$$
$$\square$$

1.A.10. Is the "square root" well defined function in complex numbers?

Solution. No, it is only defined as a function with domain being non-negative real numbers and the image being the same set.

In the complex domain, for any complex number z (except zero) there are two complex numbers such that their square is equal z. Both can be called square root and they differ by sign (square root of -1 is according to this definition i as well as $-i$). $\qquad \square$

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 $b < 0$, 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 $z = a + i b$, 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 $\operatorname{Re}(z) = a$, $\operatorname{Im}(z) = b$. It should be noted that if $z = a + i b$ and $w = c + i d$ 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

$$(a + i b) + (c + i d) = (a + c) + i (b + d),$$
$$(a + i b) \cdot (c + i d) = (ac - bd) + i (bc + ad).$$

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 $0 + i 0$, one is the number $1 + i 0$, both these numbers are for simplicity denoted as before, that is, 0 and 1. For non-zero $z = a + i b$ we easily check that $z^{-1} = (a^2 + b^2)^{-1}(a - i b)$. 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 $z = x + i y$ 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 $z \mapsto \bar{z} = x - i y$ the *complex conjugation*.

Let us now consider complex numbers of the form $z = \cos \varphi + i \sin \varphi$, where φ 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

1.A.11. Complex numbers are not just a tool to obtain "weird" solutions to quadratic equations. They are necessary to determine solutions to cubic equations, even if these solutions are real. How can we express solution to the cubic equation

$$x^3 + ax^2 + bx + c = 0$$

in real coefficients a, b, c? We show a method developed in sixteenth century by Ferro, Cardano, Tartaglia and possibly others. Substitute $x := t - a/3$ (to remove the quadratic part from the equation) to obtain the equation:

$$t^3 + pt + q = 0,$$

where $p = b - a^2/3$ and $q = c + (2a^3 - 9ab)/27$. Now introduce unknowns u, v satisfying the conditions $u + v = t$ and $3uv + p = 0$. Substitute the first condition into the previous equation to obtain

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

Now 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 $s = u^3$. Thus

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

By back substitution, we obtain

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

In the expression for u there is cube root. In order to obtain all three solutions we need to work with complex roots. The equation $x^3 = a$, $a \neq 0$, with the unknown x has exactly three solutions in the domain of complex numbers (the fundamental theorem of algebra, see (12.2.8) on page 866). All these three solutions are called cube roots of a. Therefore the expression $\sqrt[3]{a}$ has three meanings in the complex domain. If we want a single meaning for that expression, we usually consider it to be the solution with the smallest argument.

1.A.12. Show that the roots $\xi_1, \xi_2, \ldots, \xi_n$ of the equation $x^n = 1$ form the vertices of the regular n-gon in the plane of the complex numbers.

Solution. The argument of the roots is given by de Moivre theorem, namely the argument multiplied by n has to be a multiple of 2π, the absolute value has to be one, so the roots are $\xi_k = \cos(k\frac{2\pi}{n}) + i\sin(k\frac{2\pi}{n})$, $k = 1, \ldots, n$, which are indeed the vertices of a regular polygon. $\square$

numbers describe all points on the unit circle in the complex plane. Every non-zero complex number z can be then written as

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

For given $z \neq 0$, φ is unique if $0 \leq \varphi < 2\pi$. The number φ 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

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

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,

$$z\,w = |z|\,|w|(\cos(\varphi + \psi) + i\sin(\varphi + \psi))$$
$$z^n = |z|^n(\cos(n\varphi) + i\sin(n\varphi)).$$

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* $f : A \to B$ 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*.

1.A.13. Show that the roots $\xi_1, \xi_2, \ldots, \xi_n$ of the equation $x^n = 1$ satisfy

$$\sum_{i=1}^{n} \xi_i = 0.$$

Solution. Let ξ_1 be the root with the smallest positive argument. The other roots satisfy $\xi_k = \xi_1^k$ (see the previous example), thus

$$\sum_{i=1}^{n} \xi_i = \prod_{i=1}^{n} \xi_1^i = \xi_1 \frac{\xi_1^n - 1}{\xi_1 - 1} = 0,$$

where we have summed up the geometric sequence $\xi_1, \ldots, \xi_n$. $\square$

More examples about complex numbers can be found in the end of the chapter, starting at 1.G.2.

1.A.14. Solve the equation

$$x^3 + x^2 - 2x - 1 = 0.$$

Solution. This equation has no rational roots (methods to determine rational roots will be introduced later, see (**??**)). Substitution into formulas obtained in 1.A.11 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

$$\frac{14}{\sqrt[3]{3(28 - 84i\sqrt{3})}} + \frac{\sqrt[3]{28 - 84i\sqrt{3}}}{6} - \frac{1}{3} \doteq 1.247,$$

similarly for the other two (approximately -0.445 and -1.802). As noted before, we see that even if we have used complex numbers during the computation, all the solutions are real. $\square$

B. Difference equations

Difference equations (also called recurrence relations) are relations between elements of a sequence, where an element of the sequence depends on previous elements. To solve a difference equation means finding an explicit formula for n-th (that is, arbitrary) element of the sequence.

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:

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 $f : A \to B$. 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 formulas, like the formula $f(x) = 3x + 7$ 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 nth digit in the decimal expansion of $\pi \doteq 3.1415\ldots$. 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

If an element of the sequence is determined only by the previous element, we call it a first order difference equation. This is a common real world problem, for instance when we want to find out how long will repayment of a loan take for fixed monthly repayment, or when we want to know how much shall we pay per month if we want to repay a loan in a fixed time.

1.B.1. Michael wants to buy a new car. The car costs €30 000. Michael wants to take out a loan and repay it with a fixed month repayment. The car company offers him a loan to buy the car with yearly interest of 6%. The repayment starts at the end of the first month of the loan. Michael would like to finish repaying the loan in three years. How much should he pay per month?

Solution. Let P denote the sum Michael has to pay per month. After the first month Michael repays P, part of it is a repayment of the loan, part of it pays the interest. Let d_k stand for the loan after k months and write $C = 30\,000$ for the price of the car, and $u = \frac{0.06}{12}$ for the monthly interest rate. We know $d_0 = C = 30\,000$ and after the first month there is

$$d_1 = C - P + u \cdot C.$$

In general, after the k-th month we have

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

Using the relation (1) from paragraph 1.2.3 we obtain d_k given by (we write $a = 1 + u$)

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

Repaying the loan in three years means $d_{36} = 0$, thus

$$
\begin{aligned}
P &= 30\,000\left(\frac{(1+u)^{36}u}{(1+u)^{36}-1}\right) \\
&= 30\,000\left(\frac{(12.06/12)^{36}(0.06/12)}{(12.06/12)^{36}-1}\right) \doteq 912.7\,.
\end{aligned}
$$

$\qquad\square$

Note that the recurrence relation (1) can be used for our case as long as all $y(n)$ are positive, that is, as long as Michael still has to repay something.

1.B.2. Consider the case from the previous example. For how long would Michael have to pay, if he repays €500 per month?

Solution. Setting as before $a = \left(1 + \frac{0.06}{12}\right) = 1.005$, $C = 30\,000$ the condition $d_k = 0$ gives the equation

$$a^k = \frac{\frac{P}{a-1}}{\frac{P}{a-1} - C} = \frac{200P}{200P - C}.$$

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 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. function of the form $f(x) = a_n \cdot x^n + \cdots + a_1 \cdot x + a_0$, where $a_0, \ldots, a_n$ are known scalars, x is the unknown variable whose value we can insert. $x^n = 1 \cdot x \cdots x$ 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 constant 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

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

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.

By taking logarithms of both sides, we obtain

$$k = \frac{\ln(200P) - \ln(200P - C)}{\ln a},$$

which for $P = 500$ gives approximately $k = 71.5$, thus Michael would be paying for 72 months (the last repayment would be less than €500). $\square$

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

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

Linear recurrences can naturally appear in geometric problems:

1.B.4. Suppose n lines divide the plane into regions. What is the maximum number of regions that can be formed in this way?

Solution. Let the number of regions be p_n. If there is no line in the plane, then the whole plane is one region, thus $p_0 = 1$. If there are n lines, then adding an $(n + 1)$-st line increases the number of regions by the number of regions this new line intersects. If no lines are parallel and no three lines intersect at the same point, the number of regions the $(n + 1)$-st line crosses is one plus the number of its intersections with the previous lines (the crossed area will then be divided into two, thus the total number increases by one at every crossing).

The new line has at most n intersections with the already-present n lines. The segment of the line between two intersections crosses exactly one region, thus the new line crosses at most $n + 1$ regions.
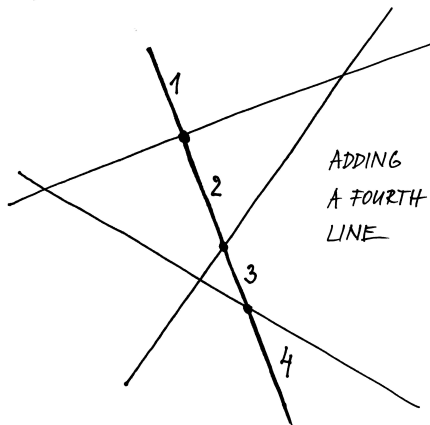
Thus we obtain the recurrence relation

$$p_{n+1} = p_n + (n + 1).$$

Such a situation can often be seen when formulating mathematical models that describe real systems in economy, biology, etc. We will observe here only a few simple examples and return to this topic in chapter 3.

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:

$$(n + 1)! = (n + 1) \cdot n!$$

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*

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

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

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

$$f(1) = af(0), \quad f(2) = af(1) = a^2 f(0), \quad \text{and so on.}$$

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.

We will prove a general result for first order equations with variable coefficients, namely:

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

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*

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

for which $p_0 = 1$. We obtain an explicit formula for p_n either by applying the formula in 1.2.2 or directly:

$$p_n = p_{n-1} + n = p_{n-2} + (n-1) + n$$

$$= p_{n-3} + (n-2) + (n-1) + n = \cdots = p_0 + \sum_{i=1}^{n} i$$

$$= 1 + \frac{n(n+1)}{2} = \frac{n^2 + n + 2}{2}.$$
□

Recurrence relations can be more complex than those of first order. We show an example of combinatorial problem, for whose solution a recurrence relation can be used.

1.B.5. How many words of length 12 that consist only of letters A and B, but do not contain a sub-word BBB, are there?

Solution. Let a_n denote the number of words of length n consisting of letters A and B but without BBB as a sub-word. Then for a_n ($n \geq 3$) the following recurrence holds

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

since the words of length n that satisfy the given condition either end with an A, or with an AB, or with an ABB. There are a_{n-1} words ending with an A (preceding the last A there can be an arbitrary word of length $n-1$ satisfying the condition). Analogously for the two remaining groups. Further, it is easily shown that $a_1 = 2$, $a_2 = 4$, and $a_3 = 7$. Using the recurrence relation we can then compute

$$a_{12} = 1705.$$

We could also derive an explicit formula for n-th element of the sequence using the theory, which we will develop in the chapter 3.
□

1.B.6. Partial difference equations. The recurrence relation in the next problem has a more complex form in comparison to the form we have dealt with in our theory. So we cannot evaluate the arbitrary member in our sequence $P_{(k,l)}$ explicitly. We can only evaluate it by a subsequent computing from previous elements. Such an equation is called partial difference equation, since the terms of the equation are indexed by two independent variables (k, l).

The score of a basketball match between the teams of Czech Republic and Russia after the first quarter is 12 : 9 for the Russian team. In how many ways could the score have developed?

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:

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

$$= \left(\prod_{i=0}^{n} a_i \right) y_0 + \sum_{j=0}^{n-1} \left(\prod_{i=j+1}^{n} a_i \right) b_j + b_n,$$

as can be seen directly by multiplying out.
□

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

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*

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

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 $(1 + a + \cdots + a^{n-1})$. The sum of this geometric progression follows from

$$1 - a^n = (1 - a)(1 + a + \cdots + a^{n-1}).$$ □

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 required the existence of the inverse element for non-zero scalars. We could not do that with integers only. Thus the last result holds for fields of scalars and we can thus use it for linear difference equations where the coefficients a, b and the initial condition $f(0) = y_0$ are rational, real or complex numbers. This last result also holds in the ring of remainder classes $\mathbb{Z}_k$ with prime k (we will define remainder classes in the paragraph 1.6.7).

It is noteworthy that the formula (1) is valid with integer coefficients and integer initial conditions. Here, we know in advance that each $f(n)$ is an integer, and the integers are a subset of rational numbers. Thus our formula necessarily gives correct integer solutions.

Observing the proof in more detail, we see that $1 - a^n$ is always divisible by $1 - a$, thus the last paragraph should not have surprised us. However it can be seen that with scalars from $\mathbb{Z}_4$ and say $a = 3$, we fail since $1 - a = 2$ is a divisor of zero and as such does not have an inverse in $\mathbb{Z}_4$.

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:
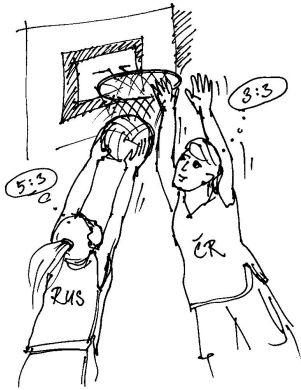
$$P_{(k,l)} = P_{(k-3,l)} + P_{(k-2,l)} + P_{(k-1,l)} + P_{(k,l-1)} + P_{(k,l-2)} + P_{(k,l-3)}.$$ Using the symmetry of the problem, $P_{(k,l)} = P_{(l,k)}$. Further, for $k \geq 3$:

$$P_{(k,2)} = P_{(k-3,2)} + P_{(k-2,2)} + P_{(k-1,2)} + P_{(k,1)} + P_{(k,0)},$$

$$P_{(k,1)} = P_{(k-3,1)} + P_{(k-2,1)} + P_{(k-1,1)} + P_{(k,0)},$$

$$P_{(k,0)} = P_{(k-3,0)} + P_{(k-2,0)} + P_{(k-1,0)},$$

which, along with the initial condition, gives $P_{(0,0)} = 1$, $P_{(1,0)} = 1$, $P_{(2,0)} = 2$, $P_{(3,0)} = 4$, $P_{(1,1)} = 2$, $P_{(2,1)} = P_{(1,1)} + P_{(0,1)} + P_{(2,0)} = 5$, $P_{(2,2)} = P_{(0,2)} + P_{(1,2)} + P_{(2,1)} + P_{(2,0)} = 14$. Hence by repeatedly using the above equations, we obtain eventually

$$P_{(12,9)} = 497178513.$$ $\square$

We will discuss recurrent formulas (difference equations) of higher order with constant coefficients in chapter 3.

C. Combinatorics

In this section we use natural numbers to describe some indivisible items located in real life space, and deal with questions as how to compute the number of their (pre)orderings, choices, and so on. In many of these problems, "common sense" is sufficient. We just need to use the rules of *product* and *sum* in the right way, as we show in the following examples:

1.C.1. Mother wants to give John and Mary five pears and six apples. In how many ways can she divide the fruits among them? (We consider the pears to be indistinguishable. We consider the apples to be indistinguishable. The possibility that one of the children gets nothing is not excluded.)

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 nth month, b_n the repayment in the nth 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.

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

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.

Solution. 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.) The six apples can be divided in seven ways. These divisions are independent. Using the rule of product, the total number is $6 \cdot 7 = 42$. □

1.C.2. Determine the number of four-digit numbers, which either start with the digit 1 and do not end with the digit 2, or that end with the digit 2 but do not start with the digit 1 (of course, the first digit must not be zero).

Solution. The set of numbers described in the statement consists of two disjoint sets. The total number is then obtained by summing the number of numbers in these two sets. In the first set there are numbers of the form "1XXY" where X is an arbitrary digit and Y is any digit except 2. Thus we can choose the second digit in ten ways, independently of that the third digit in ten ways and again independently the fourth digit in nine ways. These three choices then uniquely determine a number. By multiplication, there are $10 \cdot 10 \cdot 9 = 900$ of such numbers. Similarly in the second set we have $8 \cdot 10 \cdot 10 = 800$ numbers of the form "YXX2" (for the first digit we have only eight ways, since the number cannot start with zero and one is forbidden). By addition, the solution is $900 + 800 = 1700$ numbers. □

In the following examples we will use the notions of combinations, and permutations (possibly with repetitions).

1.C.3. During a conference, 8 speakers are scheduled. Determine the number of all possible orderings in which two given speakers do not speak one right after the other.

Solution. Denote the two given speakers by A and B. If B follows directly after the speaker A, we can consider it as a speech by a single speaker AB. The number of all orderings where B speaks directly after A is therefore 7!, 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 $8! - 2 \cdot 7!$. □

1.C.4. How many rearrangements of the letters of the word PROBLEM are there, such that

a) the letters B and R are next to each other,

b) the letters B and R are not next to each other.

Solution. a) The pair of letters B and R can be assumed to be a single indivisible "double-letter". In total we have six

$p(n)$ NON-LINEAR DEPENDENCE
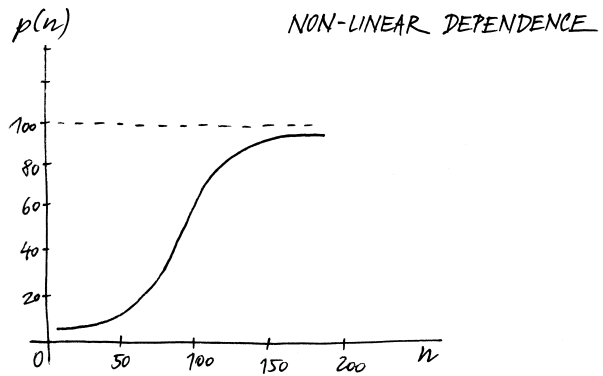
Note that the original almost exponential growth slows down later. The population size approaches the desired limit of 100 individuals. For p close to one and K much greater than r, the right side of the equation (1) is approximately $p(n)(1 + r)$. That is, the behaviour is similar to that of the Malthusian model. On the other hand, if p is almost equal to K, the right 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.[1]

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.C.1).

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

---

[1]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.

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. Thus the solution is $2 \cdot 6!$.

b) The events in b) form the complement to the part a) in the set of all rearrangements of the seven-letters. The solution is therefore $7! - 2 \cdot 6!$. □

1.C.5. In how many ways can an athlete place 10 distinct cups on 5 shelves, given that all 10 cups fit on any shelf?

Solution. Add 4 indistinguishable items, say separators, to the cups. The number of all distinct orderings of cups and separators is $14!/4!$ (the separators are indistinguishable). Each placement of cups into shelves corresponds to exactly one ordering of cups and separators. It is enough to say that the cups before the first separator in the ordering are placed in the first shelf (preserving the order), the cups between the first and the second separator in the second shelf, and so on. Thus the required number is $14!/4!$. □

1.C.6. Determine the number of four-digit numbers with exactly two distinct digits. (Recall that the first digit must not be 0.)

Solution. *First solution.* If 0 is one of the digits, then there are 9 choices for the other digit, which must also be the first digit. There are three numbers with a single 0, three numbers with two 0's, and just one number with three 0's. Thus there are 9(3+3+1)=63 numbers which contain the digit 0. Otherwise, choose the first digit for which there are 9 choices. There are then 8 choices for the other digit and 3+3+1 numbers for each choice, making $9 \cdot 8 \cdot (3+3+1) = 504$ numbers which do not contain the digit 0. The solution is 504+63=567 numbers.

Second solution. The two distinct digits used for the number can be chosen in $\binom{10}{2}$ ways. From the two chosen digits we can compose $2^4 - 2$ distinct four-digit numbers (we subtract the 2 for the two four digit numbers which use only one of the chosen digits). In total we have $\binom{10}{2}(2^4 - 2) = 630$ numbers. But in this way, we have also computed the numbers that start with zero. Of these there are $\binom{9}{1}(2^3 - 1) = 63$. Thus the solution is $630 - 63 = 567$ numbers. □

1.C.7. There are 677 people at a concert. Do some of them have the same (ordered) pair of name initials?

Solution. There are 26 letters in the alphabet. Thus the number of all possible name initials are $26^2 = 676$. Thus at least two people have the same initials. □

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 $S = \{1, \ldots, n\}$ 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.

NUMBER OF PERMUTATIONS

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

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

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 $1 \le k \le n$. 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, $0 \le k \le n$, when order does not matter.

For $k \ge 1$, 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 $c(n,k)$ of combinations of k-th degree among n elements, where $0 \le k \le n$, is*

$$(1)$$
$$c(n,k) = \binom{n}{k} = \frac{n(n-1) \ldots (n-k+1)}{k(k-1) \ldots 1} = \frac{n!}{(n-k)!k!}.$$

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

1.C.8. 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?

Solution. Each pair of players shakes hands at the introduction. The number of handshakes is then 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). Thus the required number is $6^2 = 36$. □

1.C.9. In how many ways can five people be seated in a car for five people, if only two of them have a driving licence? In how many ways can 20 passengers and two drivers be seated in a bus for 25 people?

Solution. 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 $2.4! = 48$ ways. 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, $\binom{24}{21}$. Among these seats the people can be seated in 21! ways. The solution is $2 \cdot \binom{24}{21} 21! = \frac{24!}{3}$ ways. □

1.C.10. Determine the number of distinct arrangements which can arise by permuting the letters in each individual word in the sentence "Pull up if I pull up" (the arising arrangements and words do not have to make any sense).

Solution. Let us first compute the number of rearrangements of letters in individual words. From the words "pull" we obtain $4!/2$ distinct anagrams (permutation with repetition $P(1,1,2)$), similarly "up" and "if" yields two. Therefore, using the rule of product, we have $\frac{4!}{2} \cdot 2 \cdot 2 \cdot 1 \cdot \frac{4!}{2} \cdot 2 = 1152$. Notice, that if the resulting arrangement should be a palindromic one again, there would be only four possibilities. □

1.C.11. In how many ways can we insert five golf balls into five holes (into every hole one ball), if we have four identical white balls, four identical blue balls and three identical red balls?

Solution. First solve the problem in the case that we have five balls of every colour. In this case it amounts to free choice of five elements from three possibilities (there is a choice out of

k-tuple from n parentheses in the product (from these parentheses, we take a, from the others, we take b). Therefore we have

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

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

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) $\binom{n}{k} = \binom{n}{n-k}$ $\quad 0 \le k \le n$.
(2) $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$ $\quad 0 \le k \le n-1$.
(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

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

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 $n \ge 1$. We must prove the corresponding claim for $n + 1$ using the claims (2) and (3). We calculate

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

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

three colours for every hole), that is permutations with repetitions (see). We have

$$V(3,5) = 3^5.$$

Now subtract the configurations where there are either balls of one colour (there are three such), or exactly four red balls (there are $2 \cdot 5 = 10$; we first choose the colour of the non-red ball – two ways – and then the hole it is in – five ways). Thus we can do it in

$$3^5 - 3 - 10 = 230$$

ways. □

1.C.12. In how many ways can we insert into three distinct envelopes five identical 10-bills and five identical 100-bills such that no envelope stays empty?

Solution. First compute the number of insertions ignoring the non-emptiness condition. It is an example of 3-combinations with repetition from 5 elements, and since we insert the 10-bills and 100-bills independently, we have $c(7,2)^2 = \binom{7}{2}^2$ ways. Now subtract the insertions such that exactly one envelope is empty and then the insertions such that two are empty. We have $C(2,7)^2 - 3(C(1,6)^2 - 2) - 3 = \binom{7}{2}^2 - 3(6^2 - 2) - 3 = 336$. □

1.C.13. For any fixed $n \in \mathbb{N}$, determine the number of all solutions to the equation

$$x_1 + x_2 + \cdots + x_k = n$$

in the set of non-negative integers.

Solution. Every solution $(r_1, \ldots, r_k)$, $\sum_{i=1}^{k} r_i = n$ can be uniquely encoded as a sequence of separators and ones, where we first write r_1 ones, then a separator, then r_2 ones, then another separator, and so one. Such sequence then clearly contains n ones and $k - 1$ separator. Every such sequence clearly determines some solution of the given equation. Thus there are exactly that many solutions as there are sequences, that is, $\binom{n+k-1}{n}$. □

1.C.14. In how many ways could the English Premier League have finished, if we know that no two of the three teams Newcastle United, Crystal Palace and Tottenham Hotspur are "adjacent" in the final table? (There are 20 teams in the league.)

Solution. *First approach.* We use the inclusion-exclusion principle. From the number of all possible resulting tables

says that (4) holds for some n. We calculate the corresponding sum for $n+1$ using (2) and the inductive assumption. We obtain

$$\sum_{k=0}^{n+1} k \binom{n+1}{k} = \sum_{k=0}^{n+1} k \left(\binom{n}{k-1} + \binom{n}{k} \right)$$

$$= \sum_{k=-1}^{n} (k+1) \binom{n}{k} + \sum_{k=0}^{n+1} k \binom{n}{k}$$

$$= \sum_{k=0}^{n} \binom{n}{k} + \sum_{k=0}^{n} k \binom{n}{k} + \sum_{k=0}^{n} k \binom{n}{k}$$

$$= 2^n + n2^{n-1} + n2^{n-1} = (n+1)2^n.$$

This completes the inductive step and the claim is proven for all natural n. □

The second property from above allows us to write down all the binomial coefficients into the *Pascal triangle*.[2] Here, every coefficient is obtained as a sum of the two coefficients situated right "above" it:

$n = 0:$						1					
$n = 1:$					1		1				
$n = 2:$				1		2		1			
$n = 3:$			1		3		3		1		
$n = 4:$		1		4		6		4		1	
$n = 5:$	1		5		10		10		5		1

Note that in individual rows we have the coefficients of individual powers in the expression (2). For instance the last given row says

$$(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5.$$

1.3.4. Choice with repetitions. The ordering of n elements, where some of them are indistinguishable, is called a *permutation with repetitions*.

Among n given elements, suppose there are p_1 elements of the first kind, p_2 elements of the second kind, ..., p_k of the k-th kind, where $p_1 + p_2 + \cdots + p_k = n$. Then the number of permutations with repetitions of these elements is denoted as $P(p_1, \ldots, p_k)$.

We consider the orderings which differ only in the order of indistinguishable elements to be identical. Elements of the ith kind can be ordered in $p_i!$ ways, thus we have

PERMUTATIONS WITH REPETITIONS

The number of permutations with repetitions is

$$P(p_1, \ldots, p_k) = \frac{n!}{p_1! \cdots p_k!}.$$

Let S be a set with n distinct elements. We wish to select k elements, $0 \le k \le n$ from S with repetition permitted. This is called a k-permutation with repetition. Since the first selection can be done in n ways, and similarly the second can

---

[2]Although the name goes back to Blaise Pascal's treatise from 1653, such a neat triangle configuration of the numbers $c(n,k)$ were known for centuries earlier in China, India, Greece, etc.

we subtract the tables where some two of the three teams are adjacent and then add the tables where all three teams are adjacent. The number is then

$$20! - \binom{3}{2} \cdot 2! \cdot 19! + 3! \cdot 18! = 18! \cdot 16 \cdot 17.$$

Second approach. Let us consider the three teams to be "separators". The remaining teams have to be divided such that between any two separators there is at least one team. The remaining teams can be arbitrarily permuted, as can the separators. Thus we have

$$\binom{18}{3} \cdot 17! \cdot 3! = 18! \cdot 17 \cdot 16.$$

ways. □

D. Probability

We present a few simple exercises for classical probability, where we are dealing with some experiment with only a finite number of outcomes ("all cases") and we are interested in whether or not the outcome of the experiment belongs to a subset of possible outcomes ("favourable outcomes"). The probability we are trying to determine then equals the number of favourable outcomes divided by the total number of all outcomes. Classical probability can be used when we assume, or know, that each possible outcome has the same probability of happening (for instance, fair dice throwing).

1.D.1. What is the probability that the roll of a dice results in a number greater than 4?

Solution. There are six possible outcomes (the set $\{1, 2, 3, 4, 5, 6\}$). Two are favourable ($\{5, 6\}$). Thus the probability is $2/6 = 1/3$. □

1.D.2. 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?

Solution. We divide the favourable cases according to the number of men in the chosen group: there can be either two or one. 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). There are $c(8,2) \cdot c(4,3) = \binom{8}{2} \cdot \binom{4}{3}$ of groups with two men (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

also be done in n ways etc. The total number $V(n, k)$ of k-*permutations* with repetitions is n^k. Hence

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

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 $a_1, a_2, \cdots, a_n$. Suppose each element labeled a_i is selected k_i times, $0 \leq k_i \leq k$, so that $k_1 + k_2 + \cdots + k_n = k$. 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 ith box is equal to k_i, so we obtain the sequence

$$\underbrace{* \cdots *}_{k_1} \mid \underbrace{* \cdots *}_{k_2} \mid \cdots \mid \underbrace{* \cdots *}_{k_n}.$$

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. □

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

$$1, \ 2, \ 3, \ 4, \ 5, \ 6.$$

such groups). The total number of groups with five people is $c(12,5) = \binom{12}{5}$. The probability, being the quotient of the number of favourable outcomes to the total number of outcomes, is then

$$\frac{8 + \binom{4}{3}\binom{8}{2}}{\binom{12}{5}} = \frac{5}{33}.$$

□.

1.D.3. From a deck with 108 cards ($2 \times 52 + 4$ jolly jokers) we draw without returning 4 cards randomly. What is the probability that at least one of them is an ace or a joker?

Solution. We can easily determine the probability of the complementary event, that is, in the 4 drawn cards there is none of the 12 cards (8 aces and 4 jokers). This probability is given by the ratio of the number of choices of 4 cards from 96 and the number of choices of 4 cards from 108, that is, $\binom{96}{4}/\binom{108}{4}$. The complementary event thus has the probability

$$1 - \frac{\binom{96}{4}}{\binom{108}{4}} \doteq 0.380.$$

□

We give an example for which the use of classical probability is not suitable:

1.D.4. What is the probability that the reader of this exercise wins at least 25 million euro in EuroLotto during the next week?

Solution. Such a formulation is incomplete, it does not give us enough information. We present a "wrong" solution. The sample space of possible outcomes is two-element: either the reader wins or not. A favourable event is one (win), thus the probability is $1/2$. This is clearly a wrong answer. □

Remark. In the previous exercise the basic condition of the use of classical probability was violated – every elementary event must have the same probability. In fact, the elementary event has not been defined. EuroLotto has a daily draw with a jackpot of €25 000 000 for choosing 5 correct numbers $1, \ldots, 50$. There is no other way to win €25 000 000 than to win a jackpot on some of the day during the week. The elementary event would be that a single lotto card with 5 numbers wins a jackpot. Assuming that the reader submits k lotto cards every day of the week, the probability of winning at least one jackpot during the week is $\frac{7k}{\binom{50}{5}} \doteq \frac{7k}{2\,118\,760}$.

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 $1/6$".

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 Ω. 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:

<div style="text-align:center">CLASSICAL PROBABILITY</div>

Let Ω 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 Ω of all possible states of a system (e.g. results of an experiment), which we call the *sample space*.

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. The number of seatings satisfying the given condition is $2(n!)^2$. For we have two ways for choosing the positions for men (thus also for women) – either all men sit on odd-numbered places (thus the women sit on even-numbered places), or vice versa. Among these places, both men and women are seated arbitrarily. The resulting probability is thus

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

In particular, $p(2) \doteq 0.33$, $p(5) \doteq 0.0079$, $p(8) \doteq 0.00016$. $\qquad \square$

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

i) all of them leave at sixth floor,

ii) all of them leave at the same floor,

iii) each of them leaves at a different floor.

Solution. The sample space of possible events is the space of all possible ways of leaving the elevator by 5 people. There are 8^5 of them.

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 $\frac{1}{8^4}$. 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 some person leaves the elevator and then we choose the order in which they leave the chosen floors). The probability is then (see 1.3.2 and 1.3.4)

$$\frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4}{8^5} \doteq 0.205078125. \qquad \square$$

1.D.7. Randomly choose a positive integer smaller than 10^5. What is the probability that it will consist only of the digits $0, 1, 5$ and that it will be divisible by 5? (Recall that the first digit must not be 0.)

Solution. There are $10^5 - 1$ positive integers smaller than 10^5. Numbers satisfying the condition must begin with either 1 or 5, and end with 0 or 5. Thus there are $2 \cdot 3^3 \cdot 2$ five digit favourable numbers. There are $2 \cdot 3^2 \cdot 2$ four digit favourable numbers, $2 \cdot 3 \cdot 2$ three digit favourable numbers, $2 \cdot 2$ two digit

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

PROBABILITY FUNCTION

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

(1) $\qquad\qquad P(\Omega) = 1$

(2) $\qquad\qquad 0 \le P(A) \quad$ for all events A

(3) $\quad P(A \cup B) = P(A) + P(B) \quad$ 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*.
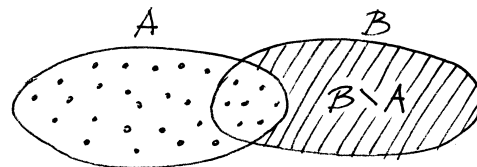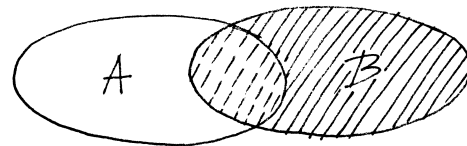
There are some further straightforward consequences of the definition for all events A, B:

(4) $\qquad\qquad P(A) = 1 - P(A^c)$

(5) $\qquad\qquad P(\emptyset) = 0$

(6) $\qquad\qquad P(A) \le 1 \quad$ for all events A

(7) $\qquad P(A) \le P(B) \quad$ whenever $A \subset B$

(8) $\qquad P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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 $P(A \cup B) = P(A \setminus B) + P(B \setminus A) + (A \cap B)$ 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 Ω 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.

favourable numbers, and one one digit favourable number. In total there are $2 \cdot (3^3 + 3^2 + 3^1 + 1) \cdot 2 + 1 = 2 \cdot (3^4 - 1) + 1 = 2 \cdot 3^4 - 1$ favourable numbers. According to classical probability, we obtain the probability as $\frac{2 \cdot 3^4 - 1}{10^5 - 1} \doteq 0.0016$. □

1.D.8. From a sack with five white and five red balls, we draw in succession three balls at random without returning the balls back to the sack. What is the probability that two of them are white and one is red?

Solution. Divide the event into a union of three disjoint events, according to in which turn we draw the red ball. The probability that the red ball is drawn as third, second, or first, respectively, is : $\frac{5}{10} \cdot \frac{4}{9} \cdot \frac{5}{8}, \frac{5}{10} \cdot \frac{5}{9} \cdot \frac{4}{8}, \frac{5}{10} \cdot \frac{5}{9} \cdot \frac{4}{8}$. In total $\frac{5}{12}$. **Another solution.** Consider the number of all possible triples of drawn balls, $\binom{10}{3}$. There are $\binom{5}{2} \cdot \binom{5}{1}$ of triples with exactly two white balls (two white balls can be drawn in $\binom{5}{2}$ ways, and one red ball can join them in five ways). The required probability is then $\binom{5}{2} \cdot \binom{5}{1} / \binom{10}{3} \doteq \frac{5}{12}$. We could forget the order, in which the balls were drawn, because every order has the same probability of being drawn. Thus there is 3! more both, favourable as well as total events and their ratio remains unchanged. □

1.D.9. From a hat where there are five white, five red and six black balls we draw balls randomly (and do not return the drawn balls back). What is the probability that the fifth drawn ball is black?

Solution. We will solve a more general problem, the probability that the i-th drawn ball is black. This probability is the same for all i, $1 \leq i \leq 16$ – we can imagine that we draw all balls one by one, and every such sequence (from the first drawn ball to the last one) consisting of five white, five red and six black has the same probability of being drawn. Thus we can use classical probability. There are $P(5,5,6) = \frac{16!}{5! \cdot 5! \cdot 6!}$ of such sequences. The number of sequences where there is a black ball on the i-th place, the rest arbitrary, equals to the number of arbitrary sequences of five white, five red and five black balls. That is, $P(5,5,5) = \frac{15!}{5!5!5!}$. Thus the probability is

$$\frac{P(5,5,5)}{P(5,5,6)} = \frac{15!}{5!5!5!} \bigg/ \frac{16!}{6!5!5!} = \frac{3}{8}.$$

□

Of course, the classical probability on the sample space Ω is an example of a probability function. The fact that the set of all events $\mathcal{A}$ is closed upon union, intersection and taking the complement has been essential in our exposition above. This will continue in all our discussion on probability in the sequel. Thus we could talk about 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 $A_1 \cup \dots A_n$. 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.

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 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.

□

1.D.10. 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 – no person receives the correct letter.

The sample space corresponds to all possible orderings of six envelopes. If we 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 ($i = 1, \ldots, 6$). These are the orderings without a fixed point. We compute the number of such orderings using the inclusion-exclusion principle. If we 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), then the resulting number d of permutations without a fixed point is

$$d = 6! - |M_1 \cup \cdots \cup M_6|$$

The number of elements in the intersection $M_{i_1} \cap \cdots \cap M_{i_k}$, $k = 1, \ldots, 6$, is $(6-k)!$ (the order of the elements $i_1, \ldots, i_k$ is fixed, the remaining $6-k$ can be ordered arbitrarily). Using the inclusion-exclusion principle we have

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

and thus for the number d we obtain the relation

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

The probability that no person receives "his" letter is then

$$\sum_{k=0}^{6} \frac{(-1)^k}{k!} = \frac{53}{144}.$$

The probability we were asked for is

$$1 - \sum_{k=0}^{6} \frac{(-1)^k}{k!} = \frac{91}{144}.$$
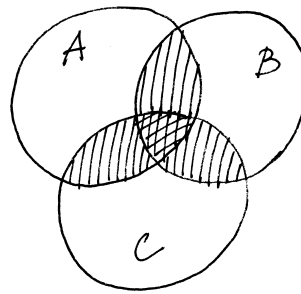
We shall now follow the same idea in order to write down the formula in the following theorem. It seems plausible that such a formula should work with proper coefficients of the sums of probabilities of intersections of more and more events among $A_1, \ldots, A_k$, at least in the case of classical probability. The reader will perhaps appreciate that a quite straightforward mathematical induction will verify the theorem in full generality.

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

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

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:

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

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

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

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

Remark. Notice that the answer does not change much with a growing number of letters. For n letters, the probability that the secretary does not assign any of them in correct order is

$$\sum_{k=0}^{n} \frac{(-1)^k}{k!} \doteq \frac{1}{e}.$$

As we will see later, the sum converges to the value $1/e$. In a similar way the exercise can be solved.

The following exercise is a simple model, which estimates the probability of death of a person in a traffic accident.

1.D.11. 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 dies in the following ten years in a traffic accident. For simplicity, assume that every person has the same "chance" of dying in traffic accident in one year and that this probability is $1200/10^7$.
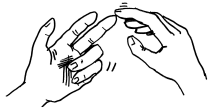
Solution. Let us first count the probability that one randomly chosen person does **not** die in ten years in a traffic accident. The probability that he/she does not die in a year is $(1 - \frac{12}{10^5})$. The probability that he/she does not die in ten years is then $(1 - \frac{12}{10^5})^{10}$. The probability that in ten years none of the given 500 people does not die is again using the product rule (the events are independent) $(1 - \frac{12}{10^5})^{5000}$. The probability of the complementary event, that is, some of the chosen people dies, is then

$$1 - \left(1 - \frac{12}{10^5}\right)^{5000} \doteq 0.4512.$$

Remark. The model we have used in the previous exercise to describe the given situation is only approximate. The complication is in the condition that every person in the sample has the same probability of dying, which is derived based on the total number of deaths per year. But the number of deaths changes yearly and even if it did not, the population changes. We show one of the possible inaccuracies by a different approach to the solution: if 1200 persons per year dies, then in ten years 12000 persons die. The probability that a certain person dies in ten years can thus be estimated by $12000/10^7$. The probability that a specific person does not die in ten years is then $(1 - \frac{12}{10^4})$ (first two members of binomial expansion of $(1 - \frac{12}{10^5})^{10}$). In total we analogously obtain the estimate

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). 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 $A_1, \ldots, A_k$. 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

$$|M \setminus (\cup_{i=1}^{k} A_i)| = |M|$$
$$+ \sum_{j=1}^{k} \left((-1)^j \sum_{1 \le i_1 < \cdots < i_j \le k} |A_{i_1} \cap \cdots \cap A_{i_j}| \right).$$

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

1.4.7. Independent events. Next, we wish to express possible dependencies among events in a given sample space Ω 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 $P(A \cap C) = \frac{1}{3} \ne \frac{1}{2} \cdot \frac{1}{2}$, $P(A \cap B) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3}$, $P(B \cap C) = \frac{1}{6} \ne \frac{2}{3} \cdot \frac{1}{2}$. Notice, that the stochastical dependance 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:

of the probability

$$1 - \left(1 - \frac{12}{10^4}\right)^{500} \doteq 0.4514.$$

We see that both estimates are very close to each other.

The effort to use mathematical knowledge for winning in various gambling games is very old. We look at a very simple example.

1.D.12. Alex has \$2500 left over from organizing a summer camp. Alex added \$50 from his savings and decided to go playing roulette. Alex bets only on colour. The probability of winning when betting on colour is $18/37$. He begins to bet \$10, and if he loses, in the next bet, he doubles the bet that he made in the previous round. (This is only if he has enough money. If not, he ends the game even if he has some money left.) If he wins, in the next round he bets again \$10. What is the probability that using this strategy he wins another \$2550? As soon as he has already won such an amount, he ends the game.

Solution. First count how many times in a row Alex can loose. If he begins with a bet of \$10, then for n bets he needs

$$10 + 20 + \cdots + 10 \cdot 2^{n-1} = 10 \cdot \left(\sum_{i=0}^{n-1} 2^i\right) = 10 \cdot (2^n - 1).$$

The number 2550 is of the form $10(2^n - 1)$ for $n = 8$. Alex can thus bet eight times in a row no matter what the result is. For nine bets he would need $10(2^9 - 1) = \$5110$ and during the game he will never have such an amount, since as soon as he has \$5100, he stops. Thus in order for him to fail, he must lose eight times in a row. The probability of losing on a single bet is $19/37$. So the probability of losing eight times in a row is $(19/37)^8$. The probability that in these eight games he wins \$10 (using his strategy) is thus $1-(19/37)^8$. In order to win \$2500, he needs to win 255 times \$10. Again using the product rule the probability of winning is

$$\left(1 - \left(\frac{19}{37}\right)^8\right)^{255} \doteq 0.29.$$

Thus the probability of winning is lower than betting everything at once on colour. $\qquad\square$

1.D.13. Individually you can try to solve the previous exercise assuming that Alex has the same strategy as before, but ends only when he has no money (if he cannot afford to double the bet when he lost the previous but still has some money, he begins again with \$10).

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

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

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

$$P(A \cap B^c) = P(A \setminus B) = P(A) - P(A \cap B) =$$
$$= P(A)(1 - P(B)) = P(A)P(B^c).$$

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 $P(A_1 \cup \cdots \cup A_k)$. In such a situation we can use the De Morgan laws for sets,

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

We obtain

$$P(A_1 \cup \cdots \cup A_k) = 1 - P(A_1^c \cap \cdots \cap A_k^c)$$
$$= 1 - (1 - P(A_1)) \cdot \ldots \cdot (1 - P(A_k))$$
$$= 1 - \prod_{j=1}^{k} (1 - P(A_j)).$$

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:

Now we consider "conditional" probability (see (1.4.8)).

1.D.14. 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?

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 Ω. Then

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

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

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

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

1.D.15. Michael has two mailboxes, one at gmail.com and the other 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% (that is, 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? (Assume that the username is always typed correctly and that making a typo cannot turn wrong password into a good 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 formula for conditional probability, is:

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

where we have used $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ since A_1 and A_2 are disjoint. We just need to determine 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

Definition. Let H be an event with non-zero probability in the sample space Ω 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" – 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 $A_1, \ldots, A_k$ satisfying $P(A_1 \cap \cdots \cap A_k) > 0$, then the theorem says

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

Notice that our condition that $P(A_1 \cap \cdots \cap A_k) > 0$ 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, thus non-zero, 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 Ω. 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 Ω with known area Ω. Events are represented by subsets $A \subset \Omega$. 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 Ω belongs to

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 $P(A_1) = \frac{1}{2} \cdot \frac{1}{20} = \frac{1}{40}$ (we multiply the probabilities, since the events are independent). Further we have (directly from the problem statement) $P(A_2) = \frac{1}{2}$. In total $P(A) = P(A_1) + P(A_2) = \frac{1}{40} + \frac{1}{2} = \frac{21}{40}$. We can evaluate

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

$\square$

The method of *geometric probability* can be used in the case that the given sample space consists of some region of a line, region, space (where we can determine (respectively) length, area, volume, ...). We assume that the probability, is equal to the ratio of the area of the subregion to the area of the sample space.

1.D.16. From Edinburgh Waverley station trains depart every hour (in the direction to Aberdeen). From Aberdeen to Edinburgh they also depart every hour. Assume that the trains move between these two stations with an uniform speed 72 km/h and are 100 meters long. The trip takes 2 hrs in either direction. The trains meet each other somewhere along the route. After visiting an Edinburgh pub, John, who lives in Aberdeen, takes the train home and falls asleep at the departure. During the trip from Edinburgh to Aberdeen he wakes up and randomly sticks his head out of the train for five seconds, on the side of the train where the trains travel in the opposite direction. What is the probability that he loses his head? (We are assuming that there are no other trains involved.)

Solution. The mutual speed of the oncoming trains is 40 metres per second, the oncoming train passes John's window for two and a half seconds. The sample space of all outcomes is thus the interval $\langle 0, 7200 \rangle$. During John's trip two trains pass by John's window in the opposite direction. Any overlap of the 2.5 seconds of the passing time interval with the 5 second time interval when John's head might be sticking out is fatal. Thus, for each train, the space of "favourable" outcomes is an interval of length 7.5 seconds somewhere in the sample space. For two trains, it is double this amount. Thus the probability of losing the head is $15/7200 \doteq 0.002$. $\square$

1.D.17. 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

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 Ω with vertex points $[0, 0]$, $[0, 1]$, $[1, 1]$ (see the diagram).
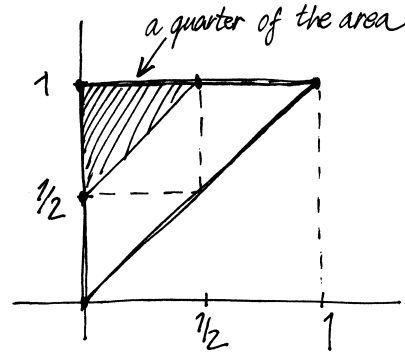
We can imagine this as a description of a problem where a very tired guest at a party tries to divide a sausage with two cuts into three pieces for himself and his two friends. What is the probability that the middle part will be at least half of the sausage?

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 ℓ such that the probability of choosing an interval (a, b) of length at least ℓ 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 Ω to be the unit square with vertices at $[0, 0]$, $[1, 0]$, $[0, 1]$, and $[1, 1]$. Let A be the intersection of Ω with the unit disk centred at the origin. Then 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!)

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. What is the probability that the buses meet, assuming they use the same route?

Solution. The sample space is a square 12×12. 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. Its area in total is 49, so the area of the "favourable part" is $144 - 49 = 95$. The probability is $p = \frac{95}{144} \doteq 0.66$.
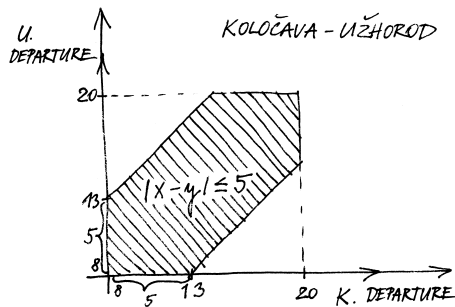
1.D.18. A rod of length two meters is randomly divided into three parts. Determine the probability that at least one part is at most 20 cm long.

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 m.
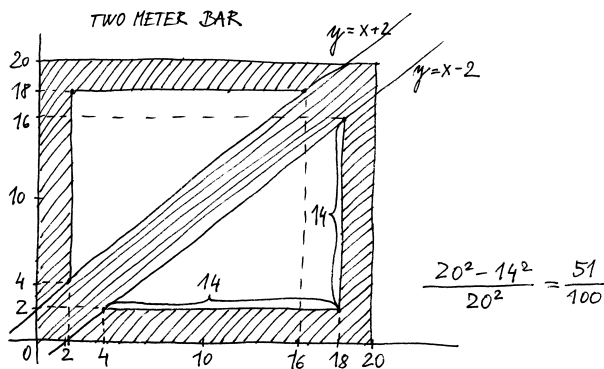
If we place the square C so that its two sides lie on axes in the plane, then the condition that at least one part is at most 20 cm determines in the square a subregion O:

$$O = \{(x, y) \in C \mid x \leq 20 \vee x \geq 180 \vee y \leq 20 \vee y \geq 180$$
$$\vee \, |x - y| \leq 20\}.$$

Of course, the well-known formula for the area of a circle with radius r is πr^2, where $\pi = 3.14159\ldots$. 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. 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.

Our tools will be mappings. We will consider only mappings 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 $a\,v$ 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

$$a\,(u + v) = a\,u + a\,v, \quad (a + b)u = a\,u + b\,u, \quad a(b\,u) = (ab)u.$$

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 combinations of vectors* e_1 and e_2. The pair of vectors $\underline{e} = (e_1, e_2)$ is called a *basis* of the vector space $\mathbb{R}^2$.

As we observe, this subregion has area $\frac{51}{100}$ times the area of the square. $\square$

E. Plane geometry

Let us start with several standard problems related to lines in plane:

1.E.1. Write down the general equation of the line $p : x = 2 - t, y = 1 + 3t, t \in \mathbb{R}$.

Solution. By eliminating t, the solution is $3x + y - 7 = 0$. $\square$

1.E.2. We are given a line

$$p : \ [2,0] + t(3,2), \ t \in \mathbb{R}.$$

Determine the general equation of this line. Determine its intersection with the line

$$q : \ [-1,2] + s(1,3), \ s \in \mathbb{R}.$$

Solution. The coordinates of the points on the first line are given by the parametric equations as $x = 2 + 3t$ and $y = 0 + 2t$. By eliminating t from the equations we obtain the equation:

$$2x - 3y - 4 = 0.$$

We obtain the intersection of p with the line q by substituting the points of q in parametric form into the equation for p:

$$2(-1 + s) - 3(2 + 3s) - 4 = 0.$$

Here we obtain $s = -12/7$ and from the parametric equation of q we obtain the coordinates of the intersection P:

$$P = \left[-\frac{19}{7}, -\frac{22}{7} \right].$$

$\square$

1.E.3. Determine the intersection of the lines

$$p : x + y - 4 = 0, \quad q : x = -1 + 2t, y = 2 + t, \ t \in \mathbb{R}.$$

Solution. Eliminate t to obtain $q : x - 2y = -5$. Then solve for x and y. The intersection has coordinates $x = 1, y = 3$. $\square$

If we choose two non-zero vectors u, v such that neither of them is a multiple of the other, then they too form a basis of $\mathbb{R}^2$.

LINEAR COMBINATION
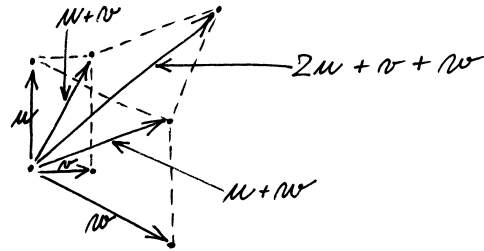
These operations are easy to imagine if we consider the vectors v to be arrows starting at the origin $0 = (0,0)$ and ending at the position (x, y) in the plane.

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.

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 usage we will always strictly 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 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.

1.E.4. Find the equation of the line p, which goes through the point $[2, 3]$ and is parallel with the line $x - 3y + 2 = 0$. Find a parametric equation of the line q which goes through the points $[1, 3]$ and $[-2, 1]$.

Solution. Every line parallel to the line $x - 3y + 2 = 0$ is given by the equation

$$x - 3y + c = 0$$

for some $c \in \mathbb{R}$. Since the line q goes through the point $[2, 3]$, $c = 7$ by putting $x = 2$ and $y = 3$. We can immediately give a parametric equation of the line q

$$q : [1, 3] + t \, (1 - (-2), 3 - 1) = [1, 3] + t \, (3, 2), \; t \in \mathbb{R}.$$

$\square$

1.E.5. Consider the following five lines. Determine if any two of the lines are parallel to each other.

$$p_1 : 2x + 3y - 4 = 0, \quad p_2 : x - y + 3 = 0,$$

$$p_3 : -2x + 2y = -6, \quad p_4 : -x - \frac{3}{2} y + 2 = 0,$$

$$p_5 : x = 2 + t, \, y = -2 - t, \; t \in \mathbb{R}$$

Solution. It is clear that

$$-2 \cdot \left(-x - \tfrac{3}{2} y + 2\right) = 2x + 3y - 4.$$

Thus p_1 and p_4 describe the same line. p_2 can be rewritten as $-2x + 2y - 6 = 0$, thus the lines p_2 and p_3 are parallel and distinct. By eliminating t, the line p_5 has an equation $x + y = 0$, which is not parallel to any other line. $\square$

1.E.6. Determine the line p which is perpendicular to the line $q : 6x - 7y + 13 = 0$ and which goes through the point $[-6, 7]$.

Solution. Since the direction vector of p is perpendicular to q, we can write the result immediately as

$$p : x = -6 + 6t, \, y = 7 - 7t, \; t \in \mathbb{R}.$$

$\square$

1.E.7. Give an example of numbers $a, b \in \mathbb{R}$, such that the vector u is a normal to AB where $A = [1, 2]$, $B = [2b, b]$, $u = (a - b, 3)$.

Solution. The direction of AB is $(2b - 1, b - 2)$ (this vector is always nonzero), and therefore the vector $(2 - b, 2b - 1)$ is normal to AB. Setting

$$2 - b = a - b, \quad 2b - 1 = 3,$$

we obtain $a = b = 2$. $\square$
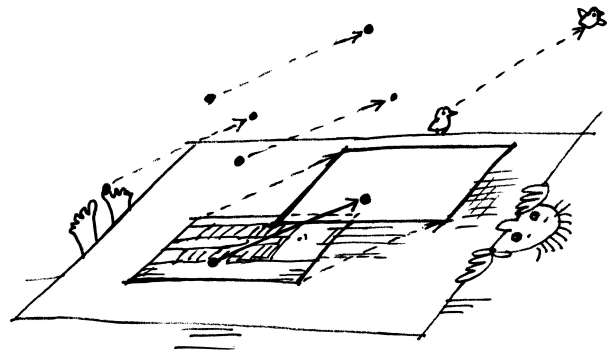
COORDINATES IN THE PLANE $\mathbb{R}^2$

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 $\mathbf{e}_1 = \overrightarrow{OE_1} = (1, 0)$. Choose any other point E_2 so that O, E_1, E_2 are distinct and not collinear. This defines the vector $\mathbf{e}_2 = \overrightarrow{OE_2} = (0, 1)$.

Then every point $P = (a, b)$ in the plane can be described uniquely as $P = O + a\mathbf{e}_1 + b\mathbf{e}_2$ for real a, b, or in vector notation, $\overrightarrow{OP} = a\mathbf{e}_1 + b\mathbf{e}_2$.

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 nonzero 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

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

1.E.8. 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 coordinates of the intersection.

Solution. Eliminating y yields $(4x - 2y - 10) + (x + 2y - 5) = 0$, from which $x = 3$, and hence $y = 1$. Hence $[3, 1]$ is the (unique) intersection and the lines are not parallel. $\square$

1.E.9. Planar soccer player shoots a ball from the point $F = [1, 0]$ in the direction $(3, 4)$ 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 $\vec{AB}$ has the parametrization $[23, 36] + u(3, -6)$, where $B = [23, 36] + 1 \cdot (3, -6)$. The intersection of these lines is given by equations $1 + 3t = 23 + 3u$ and $4t = 36 - 6u$, with the solution $t = 8$, $u = 2/3$. As $0 < 2/3 < 1$ the intersection is in the segment AB. The ball hits the goal.

Another solution. It is sufficient to consider only the slopes of the vectors $\vec{FA}$, $(3, 4)$, $\vec{FB}$. Since $\frac{36}{22} > \frac{4}{3} > \frac{30}{25}$, the player scores. $\square$

1.E.10. Consider the plane $\mathbb{R}^2$ with the standard coordinate system. A laser ray is sent from the origin $[0, 0]$ in the direction $(3, 1)$. It hits the mirror line p given by the equation

$$p : [4, 3] + t(-2, 1)$$

and then it is reflected (the angle of incidence equals the angle of reflection). At which points does the ray meet the line q, given by

$$q : [7, -10] + t(-1, 6)?$$

Solution. In principle, there could be none, one or two intersections of a ray with a line. First, we inspect possible intersection of the line q with the ray, before the ray touches the mirror line p. In the standard way, we find the intersection of the line q with the line of the initial movement of the ray: $[0, 0] + t(3, 1)$. The intersection point is $[0, 0] + \frac{91}{57}(3, 1) = [\frac{91}{19}, \frac{91}{57}]$. The ray meets the mirror at the point $[6, 2]$, that is $[0, 0] + 2(3, 1)$. As $0 < \frac{91}{57} < 2$ we conclude, that the ray meets the line q before the reflection point.

Let us concentrate now on the rebound ray. The angle between the line p and the direction of the ray can be calculated

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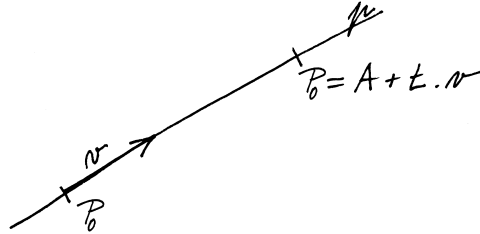
In the chosen coordinates, the point $P(t) = [x(t), y(t)]$ is given as

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

We can eliminate t from these two equations to obtain

$$-v_2 x + v_1 y = -v_2 x_0 + v_1 y_0.$$

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.

The general equation of the line in the plane is

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

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

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

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 the vector (a, b) is perpendicular to the direction of the line.

GRAPH OF A FUNCTION $f(x, y)$
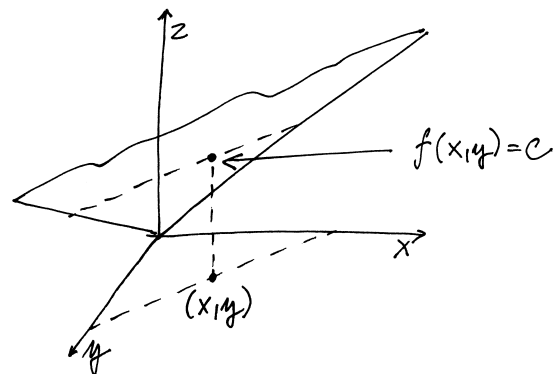
using 1.5.7 as

$$\cos\varphi = \frac{(-2,1)\cdot(3,1)}{\sqrt{5}\sqrt{10}} = -\frac{\sqrt{2}}{2},$$

therefore $\varphi = -45°$. The rebounded ray is thus perpendicular to the entering ray and its direction is $(1,-3)$. (Be careful with the orientation! The vector of the direction can also be obtained via reflection (axial symmetry) of the vector perpendicular to the line p.)

The ray meets the mirror at the point $[6,2]$, thus the reflected ray has the equation

$$[6,2] + t(1,-3),\ t \geq 0.$$

The intersection of the line given by the rebounded ray with the line q is at the point $[4,8]$. This point lies on the opposite side of the line p to both the incident and reflected rays. ($t = -2$). Thus the rebound ray does not meet the line q.

All together, there is one intersection of the ray with the line q, namely $[\frac{91}{19}, \frac{91}{57}]$. □

Remark. The reflection of a ray in three-dimensional space is studied in the exercise 3.F.4.

1.E.11. A line segment of length 1 started moving at noon with a constant speed of 1 meter per second in the direction $(3,2)$ from the point $[-2,0]$. Another line segment of length of 1 has started moving also at noon from the point $[5,-2]$ in the direction $(-1,1)$, but with double speed. Will they collide? (Segments are oriented in direction of their movements.)

Solution. Lines along which the segments are moving can be described parametrically:

$$p\ :\ [-2,0] + r(3,2),$$
$$q\ :\ [5,-2] + s(-1,1).$$

The equation of the line p is

$$2x - 3y + 4 = 0.$$

Substituting the parametric equation of the line q yields the intersection point $P = [1,2]$.

Now we choose a single parameter t for both lines so that the corresponding point on p and on q respectively, describes the position of the initial point of the first and second line segment respectively at the time t. At time 0 the initial point of the the first line segment is at $[-2,0]$, the second is at $[5,-2]$. During time t (measured in seconds) the first segments travels t units of length in the direction $(3,2)$, the second segments travels $2t$ units of length in the direction $(-1,1)$. Thus

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)
$$ax + by = r$$
$$cx + dy = s.$$

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.

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 $F\colon \mathbb{R}^2 \to \mathbb{R}^2$. 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 $f : \mathbb{R}^2 \to \mathbb{R}$ 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 2×2) 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 $\mathbf{v}$ and $\mathbf{w}$ 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

the corresponding parameterisations are:

$$p \quad : \quad [-2,0] + t\frac{(3,2)}{\sqrt{3^2+2^2}},$$

$$q \quad : \quad [5,-2] + 2t\frac{(-1,1)}{\sqrt{(-1)^2+1^2}}.$$

The initial point of the first segment enters the point $[1,2]$ at time $t_1 = \sqrt{13}$ s, the initial point of the second segment at time $t_2 = \sqrt{2}$ s – more than a half second sooner. At the time $t_2 + \frac{1}{2} = \sqrt{2} + \frac{1}{2} < t_1$ the ending point of the second segment moves away from P. Thus when the initial point of the first segment enters the point P, the ending point of the second segment is already away and the segments do not collide. □

We return for a while to complex numbers. The complex plane is basically a "normal" plane, where we have something extra:

1.E.12. Interpret multiplication by the imaginary unit i and complex conjugation as geometrical transformations in the plane.

Solution. The imaginary unit i corresponds to the point $(0,1)$. Notice that multiplying any number $z = a + ib$ by the imaginary unit i gives the result

$$i \cdot (a + ib) = -b + ia.$$

Under the interpretation in the plane, this is a rotation around the origin of the segment joining the origin to the point z through a right angle counterclockwise (cf. 1.1.4).

Taking the complex conjugate is a reflection through the axis of real numbers:

$$z = (a + ib) \mapsto (a - ib) = \bar{z}.$$

□

1.E.13. 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 $\mathbb{R}^2$.

Solution. If we view the plane $\mathbb{R}^2$ as the Gauss plane (of complex numbers), then 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 $\pi/2$. □

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

We call this the scalar product of vectors $\mathbf{v}$ and $\mathbf{w}$.

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 matrix C of correct size. In a similar way, distributivity also holds:

$$A \cdot (B + C) = A \cdot B + A \cdot C,$$

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

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

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

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

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

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* 2×2 *matrices* A. *The (vector) equation* $A \cdot v = u$ *has a unique solution for* 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

Next, we shall exercise the matrix calculus in the plane. We refer to 1.5.4 for the basic concepts. First we experience the operations of addition and multiplication on matrices, then we come to geometric tasks.
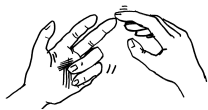
1.E.14. Simplify $(A - B)^T \cdot 2C \cdot u$, where

$$A = \begin{pmatrix} 0 & 5 \\ -2 & 2 \end{pmatrix}, \; B = \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix},$$
$$C = \begin{pmatrix} 2 & -2 \\ 4 & 5 \end{pmatrix}, \; u = \begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

Solution. By substituting in

$$A - B = \begin{pmatrix} -2 & 5 \\ -1 & 1 \end{pmatrix}, \; (A - B)^T = \begin{pmatrix} -2 & -1 \\ 5 & 1 \end{pmatrix},$$
$$2C = \begin{pmatrix} 4 & -4 \\ 8 & 10 \end{pmatrix}$$

and by matrix multiplication we obtain

$$(A - B)^T \cdot 2C \cdot u = \begin{pmatrix} -2 & -1 \\ 5 & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & -4 \\ 8 & 10 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$
$$= \begin{pmatrix} -52 \\ 64 \end{pmatrix}.$$

□

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

(a) $(A + B) \cdot (A - B) \neq A \cdot A - B \cdot B$;
(b) $(A + B) \cdot (A + B) \neq A \cdot A + 2A \cdot B + B \cdot B$.

Solution. For any two square matrices A and B we have

$$(A + B) \cdot (A - B) = A \cdot A - A \cdot B + B \cdot A - B \cdot B.$$

The identity

$$(A + B) \cdot (A - B) = A \cdot A - B \cdot B$$

is thus obtained if and only if $-A \cdot B + B \cdot A$ is the zero matrix, that is if and only if the matrices A and B commute. An example of such matrices are thus pairs of matrices, which do not commute (the matrix of product is changed when we change the order of multiplied matrices). We can choose for instance

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix},$$

since with this choice is

$$A \cdot B = \begin{pmatrix} 8 & 5 \\ 20 & 13 \end{pmatrix}, \quad B \cdot A = \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix}.$$

Notice that for any pair of square matrices A, B

$$(A + B) \cdot (A + B) = A \cdot A + A \cdot B + B \cdot A + B \cdot B.$$

It follows that

$$(A + B) \cdot (A + B) = A \cdot A + 2A \cdot B + B \cdot B$$

if and only if $A \cdot B = B \cdot A$, as in the first case. □

— 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 back to this in the next chapter. In particular we shall see that the equation (1) 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 ± 1.
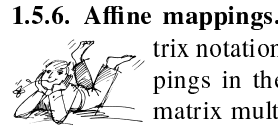
1.5.6. Affine mappings. We now investigate how the matrix notation allows us to work with simple mappings in the affine plane. We have seen that matrix multiplication defines a linear mapping. Shifting $\mathbb{R}^2$ by a fixed vector $\mathbf{w} = (r, s) \in \mathbb{R}^2$ in the affine plane can be 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.
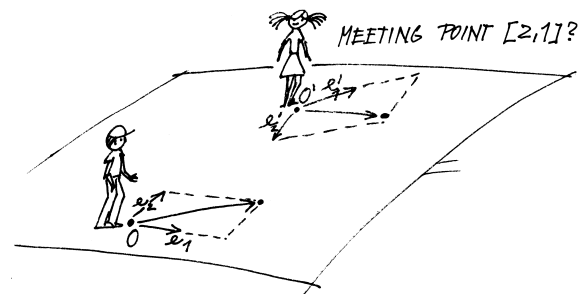
1.5.7. The distance and angle. Now we consider distance. We define the *length* of the vector $\mathbf{v} = (x, y)$ to be

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

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 $F, G : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$F : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 7x - 3y \\ -2x + 5y \end{pmatrix}, \quad x, y \in \mathbb{R},$$

$$G : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 2x + 2y - 4 \\ 4x - 9y + 3 \end{pmatrix}, \quad x, y \in \mathbb{R}$$

are linear.

Solution. For any vector $(x, y)^T \in \mathbb{R}^2$ we can express

$$F\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 7 & -3 \\ -2 & 5 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix},$$

$$G\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 2 & 2 \\ 4 & -9 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -4 \\ 3 \end{pmatrix}.$$

This implies that both mappings are affine. Recall that an affine mapping is a linear one if and only if the zero vector maps to zero. Since

$$F\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad G\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} -4 \\ 3 \end{pmatrix},$$

the mapping F is linear, the mapping G is not. Let us mention that $f(x) = Ax$ and $g(x) = Ax + b$ respectively, where $x, b \in \mathbb{R}^n$, A is a square matrix $n \times n$, are general forms of a linear and affine mappings respectively. $\square$

1.E.17. Compute the lengths of the sides of the triangle with vertices $A = [2, 2]$, $B = [3, 0]$, $C = [4, 3]$.

Solution. Using the formula for the length of a vector

$$\|u\| = \sqrt{u_1^2 + u_2^2}, \quad u = (u_1, u_2) \in \mathbb{R}^2$$

we obtain the results

$$|AB| = \|A - B\| = \sqrt{(2-3)^2 + (2-0)^2} = \sqrt{5},$$

$$|BC| = \|B - C\| = \sqrt{(3-4)^2 + (0-3)^2} = \sqrt{10},$$

$$|AC| = \|A - C\| = \sqrt{(2-4)^2 + (2-3)^2} = \sqrt{5}.$$

$\square$

1.E.18. Determine the angle between the two vectors

(a) $u = (-3, -2), v = (-2, 3)$;

(b) $u = (2, 6), v = (-3, -9)$.

Solution. The sought angle $0 \leq \varphi \leq \pi$ can of course be computed from the formula (1) in 1.5.7. But note that the vector $(-3, -2)$ can be obtained by changing the coordinates of the vector $(-2, 3)$ and multiplying one of them by the number -1. But these operations are used when we want to obtain the vector normal to a vector of direction of a given line (or vice versa). Vectors in the case (a) are thus perpendicular, that is $\varphi = \pi/2$. In the case (b), since $-3 \cdot (2, 6) = 2 \cdot (-3, -9)$, the vector u is a multiple of the vector v. If one vector is a positive multiple of another, the angle between these two is

EUCLIDEAN DISTANCE

$$\|v\| = \|P - Q\| = \sqrt{a^2 + b^2}$$
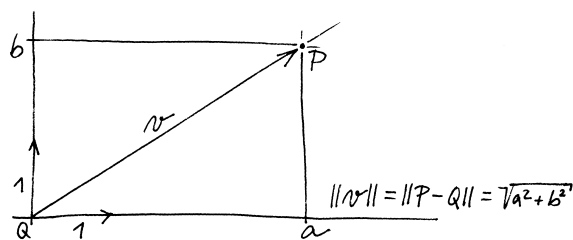
Angles are a matter of vectors rather than points in Euclidean geometry. Let $\mathbf{u}$ be a vector of length 1, at angle φ measured counter-clockwise from the vector $(1, 0)$. In coordinates, $\mathbf{u}$ is at the unit circle and has first and second coordinates $\cos \varphi, \sin \varphi$ respectively (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 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:

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

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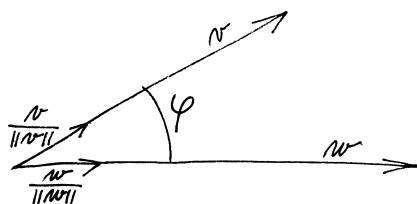
In our special case $\mathbf{v} = (1, 0)$, the more general equation gives

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

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 φ 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 is invariant with respect to rotations.

We have already seen the expression $xx' + yy'$ in the definition of the angle. We called it the *scalar product* of vectors. In the special case, when the scalar product is zero, we say that the vectors are *perpendicular*. Of course the best

zero. If it is a negative multiple, as in our case, the angle is π.

□

1.E.19. Determine the angle φ between the two diagonals A_3A_7 and A_5A_{10} of a regular dodecagon (polygon with twelve sides) $A_0A_1A_2\ldots A_{11}$.

Solution. The angle does not depend neither on the size nor on the position of the given dodecagon. Choose the dodecagon inscribed in a circle with diameter 1.

We can put A_0 to $[1,0]$ and then the vertices can be identified with the twelfth roots of 1 in the complex plane. We can write $A_k = \cos(2k\pi/12) + i\sin(2k\pi/12)$. Especially $A_3 = \cos(\pi/2) + i\sin(\pi/2) = i \sim [0,1]$, $A_5 = \cos(5\pi/6) + i\sin(5\pi/6) = -\frac{\sqrt{3}}{2} + \frac{1}{2}i \sim [-\frac{\sqrt{3}}{2}, \frac{1}{2}]$, $A_7 = \cos(7\pi/6) + i\sin(7\pi/6) = -\frac{\sqrt{3}}{2} - \frac{1}{2}i \sim [-\frac{\sqrt{3}}{2}, -\frac{1}{2}]$, and $A_{10} = \cos(5\pi/3) = i\sin(5\pi/3) = 1/2 - i\frac{\sqrt{3}}{2} \sim [1/2, -\frac{\sqrt{3}}{2}]$.

Using the formula (1) in 1.5.7 we finish the computation:

$$\cos\varphi = \frac{1}{2\sqrt{2+\sqrt{3}}},$$

that is $\varphi = 75°$.

Alternative solution. This problem can be solved via method of synthetic geometry only. Denote the centre of the regular dodecagon by S and the intersection of the diagonals A_3A_7 and A_5A_{10} by T. Now $|\angle A_7A_5A_{10}| = 45°$ (this is the inscribed angle which corresponds to the central angle A_7SA_{10}, which is a right angle), furthermore $|\angle A_5A_7A_3| = 30°$ (again the inscribed angle corresponding to the central angle A_5SA_3, which is 60°). Thus the angle A_5TA_7 is then equal to a complement of the aforementioned angles to 180°, that is 105°. The deviation we are looking for is then $180° - 105° = 75°$.

□

1.E.20. Consider a regular hexagon $ABCDEF$ with vertices labeled in the positive direction, centre at the point $S = [1,0]$ and the vertex A at $[0,2]$. Determine the coordinates of the vertex C.

Solution. The coordinates of the vertex C can be obtained by rotating the point A around the centre S of the hexagon through the angle 120° in the positive direction:

$$
\begin{aligned}
C &= \begin{pmatrix} \cos 120° & -\sin 120° \\ \sin 120° & \cos 120° \end{pmatrix}(A-S) + S \\
&= \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}\begin{pmatrix} -1 \\ 2 \end{pmatrix} + [1,0] \\
&= \left[\tfrac{3}{2} - \sqrt{3}, -1 - \tfrac{\sqrt{3}}{2}\right].
\end{aligned}
$$

□

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 the angle φ is always between 0 and π.

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 that the basis vectors $\mathbf{e}_1 = \overrightarrow{OE_1}$ and $\mathbf{e}_2 = \overrightarrow{OE_2}$ perpendicular and of length one. Such 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. The matrix of any given mapping $F : \mathbb{R}^2 \to \mathbb{R}^2$ is easy to guess. If the result of applying the mapping is 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
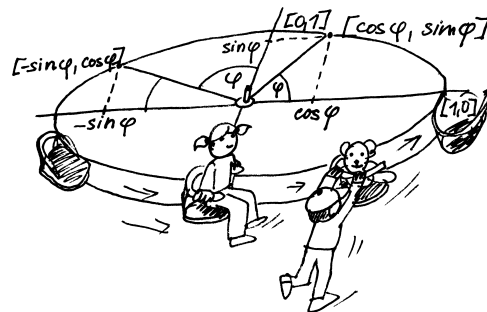
We can see from the picture that the columns of the matrix corresponding to rotating counter-clockwise through the angle ψ are computed as follows:

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

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

ROTATION MATRIX

Rotation through a given angle ψ in the positive direction about the origin is given by the matrix R_ψ:

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

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

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

1.E.21. An equilateral triangle with vertices $[1, 0]$ and $[0, 1]$ lies entirely in the first quadrant. Find the coordinates of its third vertex.

Solution. The third coordinate is $[\frac{1}{2} + \frac{\sqrt{3}}{2}, \frac{1}{2} + \frac{\sqrt{3}}{2}]$ (we are rotating the point $[1, 0]$ through $60°$ around $[0, 1]$ in the positive direction). $\qquad\square$

1.E.22. An equilateral triangle has vertices at $A = [1, 1]$ and $B = [2, 3]$. Its other vertex lies in the same half-plane as the point $S = [0, 0]$. The triangle is rotated by $60°$ in the positive direction around the point S, to produce a new triangle. Determine the coordinates of the vertices of the new triangle.

Solution. The points we are looking for have coordinates $[-\frac{3}{2}\sqrt{3}, \sqrt{3} - \frac{1}{2}], [\frac{1}{2} - \frac{1}{2}\sqrt{3}, \frac{1}{2}\sqrt{3} + \frac{1}{2}], [1 - \frac{3}{2}\sqrt{3}, \sqrt{3} + \frac{3}{2}]$. $\qquad\square$

1.E.23. Find two matrices A such that
$$A^2 = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}.$$

Hint: which geometric transformation in the plane is given by the matrix A^2?

Solution. A^2 is the matrix of rotation through $60°$ in the positive direction, thus the matrices we are looking for are
$$A = \pm \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix},$$
which are the matrices of rotation through $30°$ or through $210°$. $\qquad\square$

1.E.24. Reflection. Find the matrix of reflection in the plane through the line $y = x$ (that is, find the matrix of the axial symmetry).

Solution. The given reflection sends x-axis into y axis and vice versa. Thus the reflection applied to a vector just transposes its coordinates, therefore the sought matrix is
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A matrix of any linear mapping in $\mathbb{R}^2$ can be computed also in standard way: it is given by images of the vectors $(1, 0)$ (first column) and $(0, 1)$ (second column). In our case the images are $(0, 1)$ and $(1, 0)$. $\qquad\square$

and similarly $\mathbf{w}' = R_\psi \cdot \mathbf{w}$ for $\mathbf{w} = (r, s)^T$, and $\mathbf{w}' = (r', s')^T$. We can check that
$$\|\mathbf{v}'\| = \|\mathbf{v}\|, \text{ and that}$$
$$x'r' + y's' = xr + ys.$$

The previous expression can be written using vectors and matrices as follows:
$$(R_\psi \cdot \mathbf{w})^T (R_\psi \cdot \mathbf{v}) = \mathbf{w}^T \mathbf{v}.$$

The transposed vector $(R_\psi \cdot \mathbf{w})^T$ equals $\mathbf{w}^T \cdot R_\psi^T$, where R_ψ^T is the so-called transpose of the matrix R_ψ. 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 derivation of a remarkable claim — the matrix F with the property that $F \cdot R_\psi = I$ (we will call such a matrix the *inverse matrix to the rotation matrix* R_ψ) is the transpose of the original matrix. This makes sense, since the inverse mapping to the rotation through the angle ψ is again a rotation, but through the angle $-\psi$. That is, the inverse matrix of R_ψ^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 a matrix. 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:
$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \mathbf{v} - \mathbf{w} \mapsto R_\psi \cdot (\mathbf{v} - \mathbf{w})$$
$$\mapsto R_\psi \cdot (\mathbf{v} - \mathbf{w}) + \mathbf{w}$$
$$= \begin{pmatrix} \cos\psi\,(x - r) - \sin\psi\,(y - s) + r \\ \sin\psi\,(x - r) + \cos\psi\,(y - s)) + s \end{pmatrix}.$$

ROTATION COMPOSED WITH A SHIFT
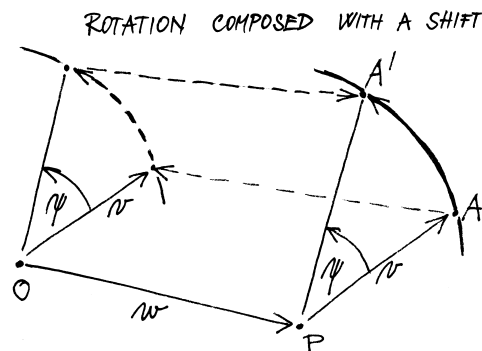
1.E.25. Determine which linear mappings from $\mathbb{R}^2$ to $\mathbb{R}^2$ are given by the following matrices (that is, describe the geometrical meaning of the matrices):

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

Solution. Let $(x, y)^T$ stand for an arbitrary real vector. For the matrix A_1 we have

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

which means that the linear mapping given by this matrix is the projection on the x axis. Similarly we can see that the matrix A_2 determines the reflection with the respect to the y axis, since

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix}.$$

The matrix A_3 can be expressed in the form

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

for $\varphi = \pi/4$, thus it gives the rotation of the plane around the origin through the angle $\pi/4$ (in the positive direction, that is counter-clockwise). $\square$

1.E.26. Show that the composition of an odd number of point reflections in the plane is again a point symmetry.

Solution. The point reflection in the plane across the point S is represented with the formula $X \mapsto S - (X - S)$, that is $X \mapsto 2S - X$. By repeated application of three point reflections across the points S, T and U respectively we obtain $X \mapsto 2S - X \mapsto 2T - (2S - X) \mapsto 2U - (2T - (2S - X)) = 2(U - T + S) - X$, that is $X \mapsto 2(U - T + S) - X$, which is a point reflection across the point $U - T + S$. Composition of any odd number of point reflections can be reduced successively to a point reflection. (In principle, this is done by mathematical induction, try to formulate it by yourself). $\square$

1.E.27. Construct a $(2n + 1)$-gon, if the middle points of all its sides are given.

Solution. We use the fact that the composition of an odd number of point reflections is again a point reflection (see the previous exercise). Denote the vertices of the $(2n + 1)$-gon we are looking for by $A_1, A_2, \ldots, A_{2n+1}$ and the middle points of the sides (starting from the middle point of $A_1 A_2$) by S_1,

1.5.9. Reflection. 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 and rotations.

We look first for a matrix Z_ψ of reflection with respect to the line through the origin and through the point $(\cos\psi, \sin\psi)$. Notice that

$$Z_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$
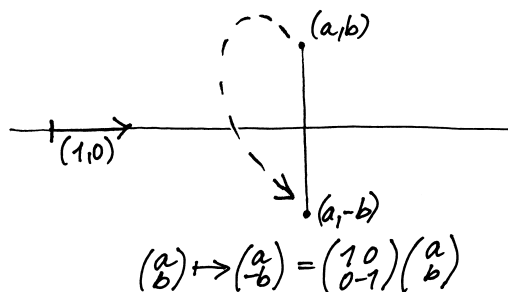
Any line going through the origin can be rotated so that it has the direction $(1, 0)$ and thus we can write general reflection matrix as

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

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_ψ.

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

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

$$= \begin{pmatrix} \cos\psi & \sin\psi \\ \sin\psi & -\cos\psi \end{pmatrix} \cdot \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix}$$

$$= \begin{pmatrix} \cos^2\psi - \sin^2\psi & 2\sin\psi\cos\psi \\ 2\sin\psi\cos\psi & -(\cos^2\psi - \sin^2\psi) \end{pmatrix}$$

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

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

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

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

37

$S_2, \ldots, S_{2n+1}$. If we carry out the point reflections across the middle points (from S_1 to S_{2n+1}), then clearly the point A_1 is a fixed point of the resulting point reflection, thus it is its centre point. In order to find it, it is enough to carry out the given point reflection with any point X of the plane. The point A_1 then lies in the middle of the line segment XX' where X' is the image of X in that point reflection. The rest of the vertices $A_2, \ldots, A_{2n+1}$ can be obtained by mapping the point A_1 in the point reflections across the points $S_1, \ldots, S_{2n+1}$. □

In the next exercises, we exploit the properties of the determinant of a matrix, cf. .

1.E.28. Determine the area of the triangle ABC, if $A = [-8, 1]$, $B = [-2, 0]$, $C = [5, 9]$.

Solution. 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 $B - A$ and the second column by the vector $C - A$, that is the determinant of the matrix

$$\begin{pmatrix} -2 - (-8) & 5 - (-8) \\ 0 - 1 & 9 - 1 \end{pmatrix}.$$

A simple calculation yields the result

$$\tfrac{1}{2} \left| (-2 - (-8)) \cdot (9 - 1) - (5 - (-8)) \cdot (0 - 1) \right| = \tfrac{61}{2}.$$

Let us add that the change of the order of the vectors leads to change in the sign of the determinant (but the absolute value is unchanged) and that the value of the determinant would not change at all if we wrote the vertices as rows (preserving the order). Moreover, the determinant formed by vectors $B - A$ and $C - A$ is always positive if the vertices A, B, C are in the anti-clockwise direction. □

1.E.29. Compute the area S of the quadrilateral given by its vertices $[1, 1]$, $[6, 1]$, $[11, 4]$, $[2, 4]$.

Solution. First, denote the vertices (in the counter-clockwise direction) as

$$A = [1, 1], \quad B = [6, 1], \quad C = [11, 4], \quad D = [2, 4].$$

If we divide the quadrilateral $ABCD$ into the triangles ABC and ACD, we can obtain its area as the sum of the areas of these two triangles, by evaluating the determinants

$$d_1 = \begin{vmatrix} 6 - 1 & 11 - 1 \\ 1 - 1 & 4 - 1 \end{vmatrix} = \begin{vmatrix} 5 & 10 \\ 0 & 3 \end{vmatrix},$$

$$d_2 = \begin{vmatrix} 11 - 1 & 2 - 1 \\ 4 - 1 & 4 - 1 \end{vmatrix} = \begin{vmatrix} 10 & 1 \\ 3 & 3 \end{vmatrix},$$
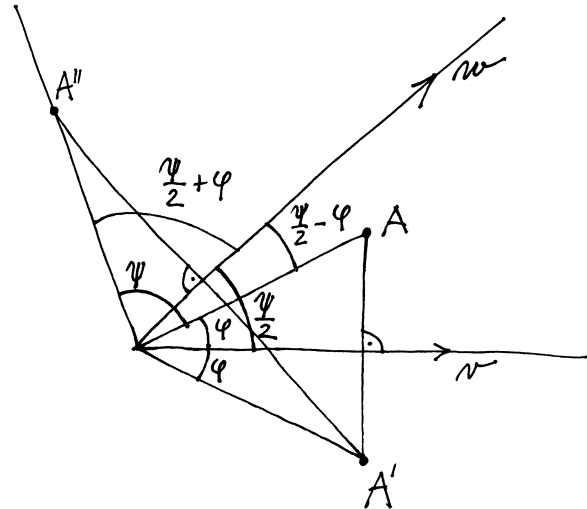
In fact we can prove the previous proposition purely by geometrical argumentation, as shown in the above picture (try to be a "synthetic geometer"). If we believe in this proof "by picture", then the above computational derivation of the proposition provides the proof of the standard double angle formulas (1).

The following is a recapitulation of previous ideas.

MAPPINGS THAT PRESERVE LENGTH

1.5.10. 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

$$x^2 + y^2 = (ax + by)^2 + (cx + dy)^2$$
$$= (a^2 + c^2)x^2 + (b^2 + d^2)y^2 + 2(ab + cd)xy.$$

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

where in the columns are these vectors $B - A$, $C - A$ (for d_1) and $C - A$, $D - A$ (for d_2). Then

$$S = \left| \frac{d_1}{2} \right| + \left| \frac{d_2}{2} \right| = \left| \frac{5 \cdot 3 - 10 \cdot 0}{2} \right| + \left| \frac{10 \cdot 3 - 1 \cdot 3}{2} \right|$$

$$= \frac{15 + 27}{2} = 21.$$

(thanks to the order of the vectors are all determinants greater than zero). Correctness of the result is easy to confirm, since the quadrilateral $ABCD$ is a trapezoid with bases of lengths 5, 9 and their distance $v = 3$. □

In the following exercises, we consider non-transparent figures (triangle, quadrangle) in the $\mathbb{R}^2$ plane. We will illustrate the power of the concept of determinant and the oriented area on practical visibility issues in the plane.

1.E.30. Visibility of the sides of a triangle. Let the triangle with the vertices $A = [5, 6]$, $B = [7, 8]$, $C = [5, 8]$ be given. Determine, which of its sides are visible from the point $P = [0, 1]$.

Solution. 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.

$$\left| \begin{matrix} B - P \\ C - P \end{matrix} \right| = \left| \begin{matrix} 7 & 7 \\ 5 & 7 \end{matrix} \right| > 0, \quad \left| \begin{matrix} C - P \\ A - P \end{matrix} \right| = \left| \begin{matrix} 5 & 7 \\ 5 & 5 \end{matrix} \right| < 0,$$

$$\left| \begin{matrix} A - P \\ B - P \end{matrix} \right| = \left| \begin{matrix} 5 & 5 \\ 7 & 7 \end{matrix} \right| = 0.$$

Not all the determinants are positive, that means P is outside the triangle. In that case, if it is left of some oriented segment (a side of the triangle), the segment is not visible from P (think this over).

Because the last determinant is zero, the points $[0, 1]$, $[5, 6]$ and $[7, 8]$ lie on a line, the side AB is thus not visible. The side BC is also not visible, unlike the side AC for which the determinant is negative. □

1.E.31. 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

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 φ. 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 φ, 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. □

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 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 does not change the area.

We can see from the statement 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 diagram that the following holds

$$\text{area } \Delta(\mathbf{v} + \mathbf{v}', \mathbf{w}) = \text{area } \Delta(\mathbf{v}, \mathbf{w}) + \text{area } \Delta(\mathbf{v}', \mathbf{w})$$

$$\text{area } \Delta(a\mathbf{v}, \mathbf{w}) = a \, \text{area } \Delta(\mathbf{v}, \mathbf{w}).$$

area of △ = 1/2 area of ▱

same area
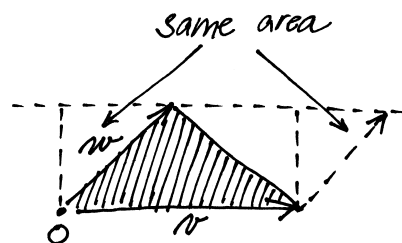
order the points without a picture; you can actually use similar procedure to what follows). First consider the side AB. It along with the point $X = [3, \pi - 2]$ determines the matrix

$$\begin{pmatrix} -2 - 3 & 2 - 3 \\ -2 - (\pi - 2) & 1 - (\pi - 2) \end{pmatrix}$$

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 oriented line $\vec{AB}$, see 1.5.12), is then determined by the sign of the determinant

$$\begin{vmatrix} -2 - 3 & 2 - 3 \\ -2 - (\pi - 2) & 1 - (\pi - 2) \end{vmatrix} = \begin{vmatrix} -5 & -1 \\ -\pi & 3 - \pi \end{vmatrix} = \\ -5 \cdot (3 - \pi) - (-1)(-\pi) < 0.$$

For the side BC we analogically obtain

$$\begin{vmatrix} 2 - 3 & 3 - 3 \\ 1 - (\pi - 2) & 3 - (\pi - 2) \end{vmatrix} = \begin{vmatrix} -1 & 0 \\ 3 - \pi & 5 - \pi \end{vmatrix} = \\ -1 \cdot (5 - \pi) - 0 < 0.$$

And for the sides CD and DA we obtain

$$\begin{vmatrix} 3 - 3 & 1 - 3 \\ 3 - (\pi - 2) & 4 - (\pi - 2) \end{vmatrix} = \begin{vmatrix} 0 & -2 \\ 5 - \pi & 6 - \pi \end{vmatrix} = \\ 0 - (-2) \cdot (5 - \pi) > 0,$$

$$\begin{vmatrix} 1 - 3 & -2 - 3 \\ 4 - (\pi - 2) & -2 - (\pi - 2) \end{vmatrix} = \begin{vmatrix} -2 & -5 \\ 6 - \pi & -\pi \end{vmatrix} = \\ -2 \cdot (-\pi) - (-5) \cdot (6 - \pi) > 0.$$

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 $\vec{XA}, \vec{XB}, \vec{XC}, \vec{XD}$ 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.32. Give 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 $[300, 1]$.

Solution. For simplifying the notation, put

$$A = [-2, -2], \quad B = [2, -11/6], \quad C = [3, 1],$$
$$D = [1, 4], \quad E = [-2, 2].$$

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

$$\begin{vmatrix} -2 - 300 & 2 - 300 \\ -2 - 1 & -\frac{11}{6} - 1 \end{vmatrix} = -302 \cdot \left(-\frac{17}{6}\right) - (-298) \cdot (-3) < 0.$$

This implies that the side can be seen from the point $[300, 1]$. $\square$

$$c = a + b$$
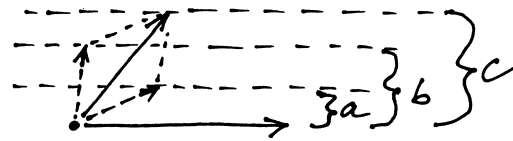
Finally we add to the formulation of our problem a condition

$$\text{area } \Delta(\mathbf{v}, \mathbf{w}) = -\text{ area } \Delta(\mathbf{w}, \mathbf{v}),$$

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 Δ 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.
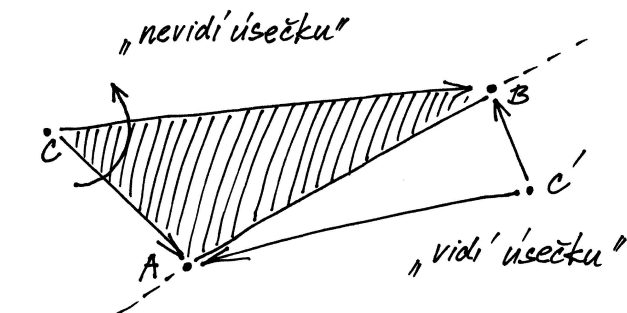
F. Relations and mappings

We conclude this chapter by considering briefly some aspects of the language of mathematics. We advise the reader to have a quick look at the definitions of the basic concepts of various relations and their properties, beginning in 1.6.1.

1.F.1. Determine whether the following relations on the set M are equivalence relations:

i) $M = \{f : \mathbb{R} \to \mathbb{R}\}$, where $f \sim g$ if $f(0) = g(0)$.

ii) $M = \{f : \mathbb{R} \to \mathbb{R}\}$, 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 always $C(n) < 10$.)

Solution.

i) We check the three properties of equivalence:

  a) Reflexivity: for any real function f, $f(0) = f(0)$.

  b) Symmetry: if $f(0) = g(0)$, then also $g(0) = f(0)$.

  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) No. The relation is not reflexive, since for instance for the function sin we have $\sin 0 \neq \sin 1$. It is not transitive.

iii) No. The relation is not reflexive (every line intersects itself). It is not transitive.

iv) Yes. The equivalence classes then correspond to unoriented directions in the plane.

v) No. The relation is not reflexive. $S(1) + S(1) = 2$. It is not transitive.

vi) Yes.

□

1.F.2. 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. If it is, describe geometrically the partitioning it determines.

We have the line segment AB and are given some point C. We calculate the oriented area of the corresponding triangle determined by the vectors $C - A$ and $B - A$. If the point C is to the left of the line segment, then with the usual positive orientation (counter-clockwise) we obtain the negative sign of the oriented area (showing the non-visibility), while the positive 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.

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

$$f : D \subseteq A \to I \subseteq B,$$
$$f(a) = b$$

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 $f : A \to B$ 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*.

Solution. From $((a,b),(a,b)) \in R$ for all $a, b \in \mathbb{R}$ it is implied that the relation is reflexive. Equally easy to see is that the relation is symmetric, since in the equality of the second coordinates we can interchange the left and right side. If $((a,b),(c,d)) \in R$ a $((c,d),(e,f)) \in R$, that is, $b = d$ and $d = f$, we easily get that the transitivity condition $((a,b),(e,f)) \in R$, that is $b = f$. The relation R is an equivalence relation, where 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). The corresponding partition then divides the plane into the lines parallel with the x axis. $\square$

1.F.3. Determine how many distinct binary relations can be defined between the 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 then correspond to subsets of this Cartesian product, and of those there are 2^{24}. $\square$

1.F.4. Give the domain D and the codomain I of the relations
$$R = \{(a,v),(b,x),(c,x),(c,u),(d,v),(f,y)\}$$
between the sets $A = \{a,b,c,d,e,f\}$ and $B = \{x,y,u,v,w\}$. Is the relation R a mapping?

Solution. Directly from the definition of the domain and the codomain of a relation we obtain
$$D = \{a,b,c,d,f\} \subset A, \quad I = \{x,y,u,v\} \subset B.$$
It is not a mapping since $(c,x),(c,u) \in R$, that is $c \in D$ has two images. $\square$

1.F.5. Determine for each of the following relations over the set $\{a,b,c,d\}$ whether it is an ordering and whether it is complete:

$R_1 = \{(a,a),(b,b),(c,c),(d,d),(b,a),(b,c),(b,d)\}$,

$R_2 = \{(a,a),(b,b),(c,c),(d,d),(d,a),(a,d)\}$,

$R_3 = \{(a,a),(b,b),(c,c),(d,d),(a,b),(b,c),(b,d)\}$,

$R_4 = \{(a,a),(b,b),(c,c),(a,b),(a,c),(a,d),(b,c),(b,d),$
$\quad (c,d)\}$,

$R_5 = \{(a,a),(b,b),(c,c),(d,d),(a,b),(a,c),(a,d),(b,c),$
$\quad (b,d),(c,d)\}$.
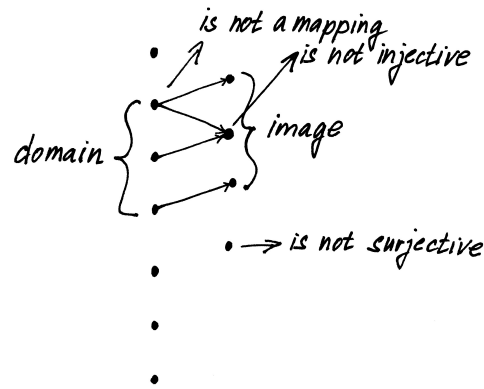
1.6.2. Composition of relations and functions. For mappings, the concept of composition is clear. Suppose we have two mappings $f : A \to B$ and $g : B \to C$. Then their *composition* $g \circ f : A \to C$ is defined as
$$(g \circ f)(a) = g(f(a)).$$
Composition can also be expressed with the notation used for a relation as
$$f \subseteq A \times B, \quad f = \{(a,f(a)); a \in A\}$$
$$g \subseteq B \times C, \quad g = \{(b,g(b)); b \in B\}$$
$$g \circ f \subseteq A \times C, \quad g \circ f = \{(a,g(f(a))); a \in A\}.$$

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*
$$\mathrm{id}_A = \{(a,a) \in A \times A; a \in A\}$$
on the set A. It is a neutral element with respect to composition with any relation that has A as its codomain or domain.
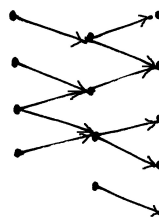
composition of relations:
points which can be reached
by a path from left to right
are in the relation

For every relation $R \subseteq A \times B$, we define the *inverse relation*
$$R^{-1} = \{(b,a); (a,b) \in R\} \subset B \times A.$$

42

Solution. 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 (it is an equivalence).

The relations R_3 and R_4 are also not an ordering, since they are not transitive (for instance $(a, b), (b, c) \in R_3, R_4$, $(a, c) \notin R_3, R_4$) and also R_4 is not reflexive ($(d, d) \notin R_d$).

The relation R_5 is a complete ordering (if we interpret $(a, b) \in R_5$ as $a \leq b$, then $a \leq b \leq c \leq d$). $\square$

1.F.6. Determine whether or not the mapping f is injective (one-to-one) or surjective (onto), when

(a) $f : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}, \quad f((x, y)) = x + y - 10x^2$;

(b) $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}, \quad f(x) = (2x, x^2 + 10)$.

Solution. In the case (a) is given a mapping which 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 case (b) it 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). $\square$

1.F.7. 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.
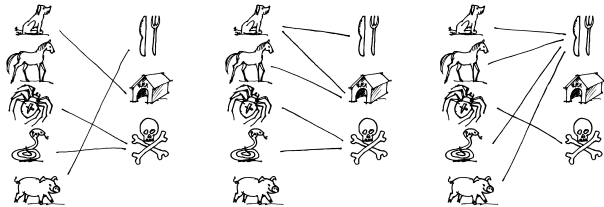
Solution. In the first figure the connection is a mapping which is surjective but not injective, because both the snake and the spider are labeled as poisonous. The second figure is not a mapping but only a relation, since the dog is labeled both as a pet and as a meal. The third connection is again a mapping. This time it is neither injective nor surjective. $\square$

1.F.8. 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. To the element 1 we can assign any of the elements a, b, c. Similarly for the element 2 we can assign any of the

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.

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 $\mathrm{id}_A \subseteq R$, 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 $R^{-1} \cap R \subseteq \mathrm{id}_A$, 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.

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.

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 ordering on $\mathbb{C}$. The absolute value is only 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

$$[a] = R_a = \{b \in A; (a, b) \in R\}.$$

elements a, b, c. Thus there are exactly 3^2 mappings of the set $\{1, 2\}$ to the set $\{a, b, c\}$. None of them can be surjective, since the set $\{a, b, c\}$ has more elements than the set $\{1, 2\}$. The mapping is injective if and only if the elements 1 and 2 are mapped to different elements. There are three possibilities for the image of 1, after 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. □

1.F.9.   Determine the number of surjective mappings of the set $\{1, 2, 3, 4\}$ to the set $\{1, 2, 3\}$.

Solution. We can determine the number 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). Thus the number of surjective mappings is

$$3^4 - \binom{3}{2}(2^4 - 2) - 3 = 36.$$

□

1.F.10.   Write down all the relations over a two-element set $\{1, 2\}$, which are symmetric but are neither reflexive nor transitive.

Solution. The reflexive relations are exactly those which contain both pairs $(1, 1), (2, 2)$. This excludes relations

$$\{(1, 1), (2, 2)\}, \quad \{(1, 1), (2, 2), (1, 2)\},$$
$$\{(1, 1), (2, 2), (2, 1)\}, \{(1, 1), (2, 2), (1, 2), (2, 1)\}.$$

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, it must by symmetry contain also the other. If it contains neither of these pairs, then it is clearly transitive. From the total number of 16 relations over a two-element set we have thus selected

$$\{(1, 2), (2, 1)\}, \qquad \{(1, 2), (2, 1), (1, 1)\},$$
$$\{(1, 2), (2, 1), (2, 2)\}.$$

It is clear that each of these 3 relations is symmetric but neither reflexive nor transitive. □

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 sometimes write $[a] = R_a$, 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) $\qquad\qquad 0 := \emptyset, \quad n + 1 := n \cup \{n\}$,

in other words

$$0 := \emptyset,\ 1 := \{0\},\ 2 := \{0, 1\}, \ldots, n + 1 := \{0, 1, \ldots, n\}.$$

This notation says that if we have already defined the numbers $0, 1, 2, \ldots n$, then the number $n + 1$ is defined as the set of all previous numbers.

We have defined the set of natural numbers $\mathbb{N}$.[3] 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 $A \to B$. 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 example 1.G.1 could be read as "most 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*.

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

$$2 = \{\emptyset, \{\emptyset\}\} \in \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\} = 4.$$

---

[3]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 developped in 19th century and there zero got a logical smallest natural number as the counterpart of the empty set.

1.F.11. Consider the set of numbers that have five digits in the binary notation and a relation such that two numbers are related whenever their digit sum has the same parity. Write down the corresponding equivalence classes.

Solution. We have two equivalence classes (of eight members): $[10000]$ $=$ $\{10000, 10011, 10101, 10110, 11001,$ $11010, 11100, 11111\}$ which corresponds to the set

$$\{16, 19, 21, 22, 25, 26, 28, 31\}$$

and $[10001]$ $=$ $\{10001, 10010, 10100, 11000, 10111,$ $11011, 11101, 11110\}$ which corresponds to the set

$$\{17, 18, 20, 24, 23, 27, 29, 30\}.$$

$\square$

1.F.12. Consider the set of numbers that have three digits in the ternary notation and a relation such that two numbers are in the relation whenever they

 i) begin with the same two digits in this notation,
 ii) end with the same two digits in this notation.

Write down the corresponding equivalence classes.

Solution.

 i) We obtain six three-element classes

$[100] = \{100, 101, 102\}$ corresponds $\{9, 10, 11\}$

$[110] = \{110, 111, 112\}$ corresponds $\{12, 13, 14\}$

$[120] = \{120, 121, 122\}$ corresponds $\{15, 16, 17\}$

$[200] = \{200, 201, 202\}$ corresponds $\{18, 19, 20\}$

$[210] = \{210, 211, 212\}$ corresponds $\{21, 22, 23\}$

$[220] = \{220, 221, 222\}$ corresponds $\{24, 25, 26\}.$

 ii) In this case we have nine two-element classes

$[100] = \{100, 200\}$ corresponds $\{9, 18\}$

$[101] = \{101, 201\}$ corresponds $\{10, 19\}$

$[102] = \{102, 202\}$ corresponds $\{11, 20\}$

$[110] = \{110, 210\}$ corresponds $\{12, 21\}$

$[111] = \{111, 211\}$ corresponds $\{13, 22\}$

$[112] = \{112, 212\}$ corresponds $\{14, 23\}$

$[120] = \{120, 220\}$ corresponds $\{15, 24\}$

$[121] = \{121, 221\}$ corresponds $\{16, 25\}$

$[122] = \{122, 222\}$ corresponds $\{17, 26\}.$

$\square$

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 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. 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.

If we define multiplication of integers similarly to the addition, 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 ± 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 $\mathbb{N}$. 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.F.13. Determine the number of equivalence relations over a set $\{1, 2, 3, 4\}$.

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

In total we have 15 different equivalences. $\square$

Remark. In general, the number of partitions of a given n-element set is given by the *Bell number* B_n, satisfying a recurrence formula

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k.$$

(devide the partitions according to the cardinality of the set, to which one fixed element belongs)

1.F.14. 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:
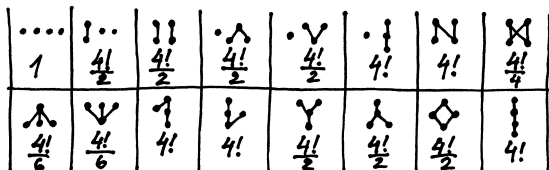
In total, there are 219 orderings over a four-element set. $\square$

There are many combinatorics problems which refer to relations. You can find some of them in the additional exercises after this chapter, starting with 1.G.71.

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 using representatives we can correctly define addition and multiplication for each $\mathbb{Z}_k$.

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 $[x]$. $\square$

G. Additional exercises for the whole chapter

1.G.1. Let t and m be positive integers. Show that the number $\sqrt[m]{t}$ is either integer or is not rational.

Solution. We shall exploit the basic divisibility rules of integers which we shall discuss in detail later in chapter 11. Show that if the number is not integer, then it cannot be rational. If $\sqrt[m]{t}$ is not integer, then there exists a prime r and integer s such that r^s divides t, r^{s+1} does not divide t (this we write as $\operatorname{ord}_r t = s$) and m does not divide s. Assume that $\sqrt[m]{t} = \frac{p}{q}$, $p, q \in \mathbb{Z}$, in other words $t \cdot q^m = p^m$. Consider $\operatorname{ord}_r L$ and $\operatorname{ord}_r R$ and their divisibility by the number m. (L and R denote the left-hand and right hand side of the equation respectively). $\square$

1.G.2. Find the algebraic form of the expressions:

    i) $\frac{2+i}{5+3i}$,

    ii) $\frac{(1+i)^2}{(\sqrt{2}+\sqrt{2}i)^{30}}$.

Solution. By expanding the fraction by $5 - 3i$ we get easily the result $\frac{13}{34} - \frac{1}{34}i$. The second task is an explicit computation resulting in $2^{-29}i$. $\square$

1.G.3. In the complex plane draw the solutions of the equations:

    i) $z = |z|$,

    ii) $|z^2 + 1| = 1$,

    iii) $\operatorname{Re} z = \operatorname{Re}(z + 1)$.

Solution. Write the solution!! $\square$

1.G.4. Mark the following sets in the complex plane:
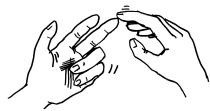
    i) $\{z \in \mathbb{C} \mid |z - 1| = |z + 1|\}$,

    ii) $\{z \in \mathbb{C} \mid 1 \le |z - i| \le 2\}$,

    iii) $\{z \in \mathbb{C} \mid \operatorname{Re}(z^2) = 1\}$,

    iv) $\{z \in \mathbb{C} \mid \operatorname{Re}(\frac{1}{z}) < \frac{1}{2}\}$.

Solution. (i) the imaginary axis,
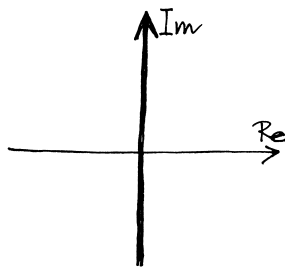
(ii) annulus aroumd i,
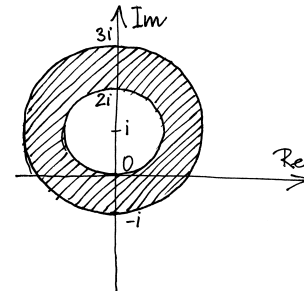
(iii) hyperbola $a^2 - b^2 = 1$,
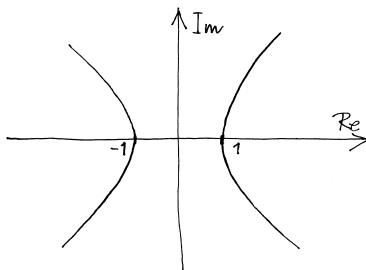
(iv) exterior of the unit disc centered at 1.

$\square$
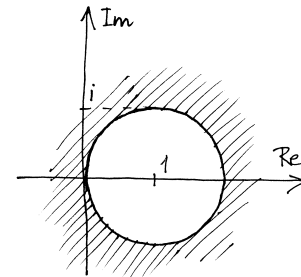
1.G.5. Consider an "assignment", which sends every real number a to a root of $x^2 + x + a = 0$. Does it give a function $\mathbb{R} \to \mathbb{C}$?

Solution. No, the prescription is not unique, there is always a choice from two numbers, except for $a = 1/4$. $\square$

1.G.6. Determine the number of ways of placing the white tower and black tower on the chessboards (of size 8×8), that are neither in the same column nor in the row.

Solution. First, we can place the white tower in any of 8^2 positions. Then we have "to our disposal" 7^2 positions in which to place the black tower. The total number of ways is $8^2 \cdot 7^2 = 3\,136$. $\square$

1.G.7. There were six men in the meeting. If all of them shook hands with each other, how many handshakes have happened?

Solution. The number of handshakes equals the number of ways of choosing an unordered tuple among 6 elements, thus the result is $c\,(6, 2) = \binom{6}{2} = 15$. $\square$

1.G.8. Determine in how many ways a four-member committee can be chosen among 15 deputies, if it is not allowed for two certain deputies to work together.

Solution. The result is

$$\binom{15}{4} - \binom{13}{2} = 1\,287.$$

It can be obtained by first calculating the number of all four-member committees and then subtracting the number of those committees where the given two deputies are chosen together (in that case, we only choose two more members among the remaining 13 deputies). $\square$

1.G.9. In how many ways can we divide 8 women and 4 men in two six-member groups (which are considered unordered) in such a way that there is at least one man in each group?

Solution. If we forget the last condition, division of 12 people in two six-member groups can be done by just choosing 6 people and put them to the first group, which can be done in $\binom{12}{6}$ ways. The groups are not distinguishable (we do not know which one is the first one), thus the total number is rather $\frac{1}{2} \cdot \binom{12}{6}$. In $\binom{8}{2}$ cases all men are in one group (we choose two women among eight to complete the group). The correct answer is thus

$$\tfrac{1}{2} \cdot \binom{12}{6} - \binom{8}{2} = 434.$$

$\square$

1.G.10. Determine the number of even four-digit numbers composed of exactly two distinct digits.

Solution. Analogously to 1.C.6, we ignore first the peculiarities of the digit zero. We obtain $\binom{5}{2}(2^4 - 2) + 5 \cdot 5(2^3 - 1)$ numbers (In the first summand, we count the numbers that consist only of even digits. In the second summand we count the number of even four-digit numbers with one digit even and one digit odd). Again we have to subtract the numbers that start with zero, of those there are $(2^3 - 1)4 + (2^2 - 1)5$. The final number is thus

$$\binom{5}{2}(2^4 - 2) + 5 \cdot 5(2^3 - 1) - (2^3 - 1)4 - (2^2 - 1)5 = 272.$$

$\square$

1.G.11. What is the number of 4-digit numbers composed of digits $1, 3, 5, 6, 7$ and 9, where no digit occurs more than once?

Solution. We have 6 distinct letters at our disposal. We ask: how many distinct ordered 4-tuples can be chosen from them? The result is $v\,(6, 4) = 6 \cdot 5 \cdot 4 \cdot 3 = 360$. $\square$

1.G.12. 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. The result is then $V(24, 5) = 24^5$. $\square$

1.G.13. 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

$$(k - 3)! \cdot (k - 2) \cdot (k - 3) \cdot (k - 4) = (k - 2)! \cdot (k - 3) \cdot (k - 4).$$ □

1.G.14. 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. 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

$$P(10, 8, 8, 6) = \frac{32!}{10! \cdot 8! \cdot 8! \cdot 6!}.$$ □

1.G.15. 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. If we 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, then we create permutations with repetitions from the elements 1, 2, 3, where 1 occurs four times, 2 three times and 3 two times. Number of such permutations is

$$P(4, 3, 2) = \frac{9!}{4! \cdot 3! \cdot 2!} = 1\,260.$$ □

1.G.16. Determine the number of ways how to divide among three people A, B and C 33 distinct coins such that A and B together have twice as many coins as C.

Solution. 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 $\binom{33}{11} \cdot 2^{22}$. □

1.G.17. In how many ways can we divide 40 identical balls among 4 boys?

Solution. 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 $\binom{43}{3} = 12\,341$. □

1.G.18. 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.19. In how many ways could the table of the first soccer league ended, if we know only that at least one of the teams Ostrava, Olomouc is in the table after the team of Brno (there are 16 teams in the league).

Solution. Let us first determine the three places where the teams of Brno, Olomouc and Ostrava ended. Those can be chosen in $c(3, 16) = \binom{16}{3}$ ways. From 6 possible orderings of these three teams on the given three places only four satisfy the given condition. After that, we can independently choose the order of the remaining 13 teams at the remaining places of the table. Using the rule of product, we have the solution

$$\binom{16}{3} \cdot 4 \cdot 13! = 13948526592000.$$ □

1.G.20. How many distinct orderings (in a row) at a picture of a volleyball team (6 players), if

i) Gouald and Bamba want to stand next to each other;

ii) Gouald and Bamba want to stand next to each other and in the middle;

iii) Gouald and Kamil do not want to stand next to each other.

Solution.

i) In this case Gouald a Bamba can be considered a single person, we just multiply then by two to determine their relative order. Thus we have $2.5! = 240$ orderings.

ii) Here it is similar except that the position of Gouald and Bamba is fixed. We have $2.4! = 48$ orderings.

iii) Probably the simplest approach is to subtract the cases where Kamil and Gouald stand next to each other (see (i)). We get $6! - 2.5! = 720 - 240 = 480$. □

1.G.21. Coin flipping. We flip a coin six times.

i) How many distinct sequences of heads and tails are there?

ii) How many sequences with exactly four heads are there?

iii) How many sequences with at least two heads are there? ○

1.G.22. How many distinct anagrams (rearrangements of letters) of the word "krakatit", such that between the letters "k" there is exactly one other letter.

Solution. In the considered anagrams there are exactly six possibilities of placement of the group two "k", since the first of the two "k" can be placed at any of the positions $1 - 6$. If we fix the spots for the two "k", then the other letters can be placed arbitrarily, that is, in $P(1, 1, 2, 2)$ ways. Using the rule of product, we have

$$6 \cdot P(1, 1, 2, 2) = \frac{6 \cdot 6!}{2 \cdot 2} = 1080.$$

□

1.G.23. How many anagrams of the word BASILICA are there, such that there are no two vowels next to each other and no two consonants next to each other?

Solution. Since there are four vowels and four consonants in the word, each such anagram is either of the type $BABABABA$ or $ABABABAB$. On the given four places we can permute vowels in $P_o(2, 2) = \frac{4!}{2!2!}$ ways and independently of that also the consonants (4! ways). Using the rule of product, the result is then $2 \cdot 4! \cdot \frac{4!}{2!2!} = 288$. □

1.G.24. In how many ways can we divide 9 girls and 6 boys into two group such that each group contains at least two boys?

Solution. We divide the boys and the girls independently: $2^9(2^5 - 7) = 12800$. □

1.G.25. Material is composed of five layers, each of them has fibres in one of the possible six directions. How many of such materials are there? How many of them have no two neighbouring layers which have fibres in the same direction?

Solution. 6^5 and $6 \cdot 5^5$. □

1.G.26. For any fixed $n \in \mathbb{N}$ determine the number of all solutions to the equation

$$x_1 + x_2 + \cdots + x_k = n$$

in the set of positive integers.

Solution. If we look for a solution in the domain of positive integers, then we note that the natural numbers $x_1, \ldots x_k$ are a solution to the equation if and only if the non-negative integers $y_i = x_i - 1$, $i = 1, \ldots, k$ are a solution to the equation

$$y_1 + y_2 + \cdots + y_k = n - k.$$

Using 1.C.13, there are $\binom{n-1}{k-1}$ of them. □

1.G.27. There are n forts on a circle ($n \geq 3$), numbered in a row with numbers $1, \ldots, n$. In one moment of time each of the forts shoots at one of its neighbours (fort 1 neighbours also with the fort n). Denote by $P(n)$ the number of all possible results of the shooting (a result of the shooting is a set of numbers of those forts that were hit, regardless of the number of hits taken). Prove that $P(n)$ and $P(n+1)$ are relatively prime.

Solution. If we denote the forts that were hit by a black dot and the unhit by a white dot, the task is equivalent to the task to determine the number of all possible colourings of n dots on a circle with black and white colour, such that no two white dots have "distance" one. For odd n this number is equal to $K(n)$ – the number of colourings with black and white, such that no two white dots are adjacent (we reorder the dots such that we start with the dot one and proceed increasingly with odd numbers, and then increasingly with even). For even n this number equals $K(n/2)^2$, the square of the colouring of $n/2$ dots on a circle such that no two white are adjacent (we colour independently the dots on even positions and on odd positions).

For $K(n)$ we easily derive a recurrent formula $K(n) = K(n-1) + K(n-2)$. Furthermore, we can easily compute that $K(2) = 3$, $K(3) = 4$, $K(4) = 7$, that is, $K(2) = F(4) - F(0)$, $K(3) = F(5) - F(1)$, $K(4) = F(6) - F(2)$, and using induction we can easily prove that $K(n) = F(n+2) - F(n-2)$, where $F(n)$ denotes the n-th member of the Fibonacci sequence ($F(0) = 0$, $F(1) = F(2) = 1$). Since $(K(2), K(3)) = 1$, we have for $n \geq 3$ similarly as in the Fibonacci sequence

$$(K(n), K(n-1)) = (K(n) - K(n-1), K(n-1)) = (K(n-2), K(n-1)) = \cdots = 1.$$

Let us now show that for every even $n = 2a$ is $P(n) = K(a)^2$ relatively prime with both $P(n+1) = K(2a+1)$ and $P(n-1) = K(2a-1)$. For this the following is enough: for $a \geq 2$ we have

$$(K(a), K(2a+1)) = (K(a), F(2)K(2a) + F(1)K(2a-1)) = (K(a), F(3)K(2a-1) + F(2)K(2a-2)) = \ldots$$
$$= (K(a), F(a+1)K(a+1) + F(a)K(a)) = (K(a), F(a+1)) = (F(a+2) - F(a-2), F(a+1))$$
$$= (F(a+2) - F(a+1) - F(a-2), F(a+1)) = (F(a) - F(a-2), F(a+1))$$
$$= (F(a-1), F(a+1)) = (F(a-1), F(a)) = 1, \text{ and}$$
$$(K(a), K(2a-1)) = (K(a), F(2)K(2a-2) + F(1)K(2a-3)) = (K(a), F(3)K(2a-3) + F(2)K(2a-4))$$
$$= \cdots = (K(a), F(a)K(a) + F(a-1)K(a-1)) = (K(a), F(a-1)) = (F(a+2) - F(a-2), F(a-1))$$
$$= (F(a+2) - F(a), F(a-1)) = (F(a+2) - F(a+1), F(a-1)) = (F(a), F(a-1)) = 1.$$

This proves the claim. □

1.G.28. How much money do I save in a building savings in five years, if I invest in it 3000 Kč monthly (at the first day of the month), the yearly interest rate is 3% and once a year I obtain a state donation of 1500 Kč (this donation comes at first of May)?

Solution. Let x_n be the amount of money at the account after n years. Then (for $n > 2$) we obtain the following recurrent formula (assuming that every month is exactly one twelfth of a year)

$$x_{n+1} = 1.03\, x_n + 36000 + 1500 + \underbrace{0.03 \cdot 3000 \left(1 + \frac{11}{12} + \cdots + \frac{1}{12}\right)}_{\text{interests from deposits this year}} + \underbrace{0.03 \cdot \frac{2}{3} \cdot 1500}_{\text{interest from the state donation credited at this year}} = 1.03\, x_n + 38115.$$

Therefore

$$x_n = 38115 \sum_{i=0}^{n-2} (1.03)^i + (1.03)^{n-1} x_1 + 1500,$$

while $x_1 = 36000 + 0.03 \cdot 3000 \left(1 + \frac{11}{12} + \cdots + \frac{1}{12}\right) = 36585$, in total

$$x_5 = 38115 \left(\frac{(1.03)^4 - 1}{0.03}\right) + (1.03)^4 \cdot 36585 + 1500 \doteq 202136.$$

□

Remark. In reality, interests are computed according to the number of days the money is on the account. You should obtain a real bank statement of a building savings, determine its interest rates and try to compute the credited interests in a year. Compare the result with the sum that was credited in reality. Compute until the numbers agree …

1.G.29. What is the maximum number of areas the plane can be divided into by n circles?

Solution. For the maximum number p_n of areas we derive a recurrent formula

$$p_{n+1} = p_n + 2n.$$

Note that the $(n+1)$-th circle intersects n previous circles in at most $2n$ points (and this can really occur)

ADDING THE THIRD CIRCLE
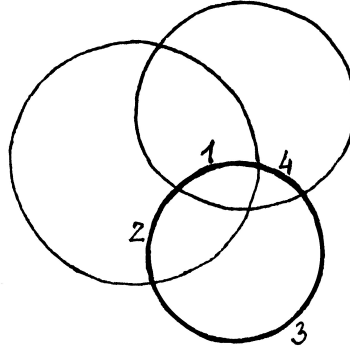
Clearly $p_1 = 2$. Thus for p_n we obtain

$$p_n = p_{n-1} + 2(n-1) = p_{n-2} + 2(n-2) + 2(n-1) = \ldots$$

$$= p_1 + \sum_{i=1}^{n-1} 2i = n^2 - n + 2.$$

$\square$

1.G.30. What is the maximum number of areas a 3-dimensional space can be divided into by n planes?

Solution. Let the number be r_n. We see that $r_0 = 1$. Similarly to the exercise (1.B.4) we consider n planes in the space, we add another plane ad we ask what is the maximum number of new areas. Again it is exactly the number of areas the new plane intersects. How many can that be? The number of areas intersected by the $(n+1)$-th plane equals to the number of areas the new $(n+1)$-th plane is divided into by the lines of intersection with the n planes that were already situated in the space. However, there are at most $1/2 \cdot (n^2 + n + 2)$ of those (according to the exercise in plane), thus we obtain the recurrent formula

$$r_{n+1} = r_n + \frac{n^2 + n + 2}{2}.$$

This equation can be again solved directly:

$$
\begin{aligned}
r_n &= r_{n-1} + \frac{(n-1)^2 + (n-1) + 2}{2} = r_{n-1} + \frac{n^2 - n + 2}{2} \\
&= r_{n-2} + \frac{(n-1)^2 - (n-1) + 2}{2} + \frac{n^2 - n + 2}{2} \\
&= r_{n-2} + \frac{n^2}{2} + \frac{(n-1)^2}{2} - \frac{n}{2} - \frac{(n-1)}{2} + 1 + 1 \\
&= r_{n-3} + \frac{n^2}{2} + \frac{(n-1)^2}{2} + \frac{(n-3)^2}{2} - \frac{n}{2} - \frac{(n-1)}{2} - \frac{(n-2)}{2} \\
&\quad + 1 + 1 + 1 \\
&= \cdots = r_0 + \frac{1}{2}\sum_{i=1}^{n} i^2 - \frac{1}{2}\sum_{i=1}^{n} i + \sum_{i=1}^{n} 1 \\
&= 1 + \frac{n(n+1)(2n+1)}{12} - \frac{n(n+1)}{4} + n =
\end{aligned}
$$

$$= \frac{n^3 + 6n + 5}{6},$$

where we have used the known relation

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

which can be easily proved by mathematical induction. □

1.G.31. What is the maximum number of areas a 3-dimensional space can be divided into by n balls? ○

1.G.32. What is the number of areas a 3-dimensional space is divided into by n mutually distinct planes which all intersect a given point?

Solution. For the number x_n of areas we derive a recurrent formula

$$x_n = x_{n-1} + 2(n-1),$$

furthermore $x_1 = 2$, that is,

$$x_n = n(n-1) + 2.$$

□

1.G.33. From a deck of 52 cards we randomly draw 16 cards. Express the probability that we choose exactly 10 red and 6 black cards.

Solution. We first realize that we don't have to care about the order of the cards. (In the resulting fraction we would obtain ordered choices by multiplying by 16! both nominator and denominator.) The number of all possible (unordered) choices of 16 cards from 52 is $\binom{52}{16}$. Similarly, the number of all choices of 10 cards from 26 is equal to $\binom{26}{10}$ and of 6 cards from 26 is $\binom{26}{6}$. Since we are choosing independently 10 cards from 26 red and 6 cards from 26 black, using the (combinatorial) rule of product we obtain the result

$$\frac{\binom{26}{10} \cdot \binom{26}{6}}{\binom{52}{16}} \doteq 0.118.$$

□

1.G.34. In a box there are 7 white, 6 yellow and 5 blue balls. We draw (without returning) 3 balls randomly. Determine the probability that exactly 2 of them are white.

Solution. In total there are $\binom{7+6+5}{3}$ ways, how to choose 3 balls. Choosing exactly two white allows $\binom{7}{2}$ choices of two white balls and simultaneously $\binom{11}{1}$ choices for the third ball. Using the rule of product is the number of ways how to choose exactly two white equal to $\binom{7}{2} \cdot \binom{11}{1}$. Thus the result is

$$\frac{\binom{7}{2} \cdot 11}{\binom{18}{3}} \doteq 0.283.$$

□

1.G.35. When throwing a dice, eleventh times in a row the result was 4. Determine the probability that the twelfth roll results in 4.

Solution. The previous results (according to our assumptions) do not influence the result of further rolls. Thus the probability is $1/6$. □

1.G.36. From a deck of 32 cards we randomly draw 6 cards. What is the probability that all of them have the same colour?

Solution. In order to obtain the result

$$\frac{4 \cdot \binom{8}{6}}{\binom{32}{6}} \doteq 1.234 \cdot 10^{-4},$$

we just first choose one of the 4 colours and realize that there are $\binom{8}{6}$ ways how to choose 6 cards from 8 cards of this colour.

□

1.G.37. Three players are given 10 cards each and two remain (from a deck of 32 cards, where 4 of them are aces). Is it more likely, that somebody receives seven, eight and nine of spades; or that two aces remain?

Solution. Since the probability that some of the players receives the three mentioned cards equals

$$3\,\frac{\binom{29}{7}}{\binom{32}{10}},$$

while the probability that two aces remain equals

$$\frac{\binom{4}{2}}{\binom{32}{2}},$$

it is more likely that some of the players receives the three mentioned cards. Let us note that proving the inequality

$$\frac{3 \cdot \binom{29}{7}}{\binom{32}{10}} > \frac{\binom{4}{2}}{\binom{32}{2}}$$

is possible by transforming both sides, where by repetitive crossing-out (after expanding the binomial coefficients according to their definition) we easily obtain $6 > 1$. □

1.G.38. We throw n dice. What is the probability that among the numbers that appeared the values 1, 3 and 6 are not present?

Solution. We can reformulate the exercise that we throw 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 $1/4$ (the result of the first roll does not influence the result of the second roll). Since the event determined by the result of a given roll and event determined by the result of another roll are always (stochastically) independent, the probability is $1/2^n$. □

1.G.39. Two friends are shooting independently of each other at one target – one shoots, then the second shoots, then the first, and so on. The probability that the first hits is 0.4, the second friend has the probability of hitting 0.3. Determine the probability P of the event that after shooting there will be exactly one hit of the target.

Solution. We determine the result by summing the probabilities of two mutually exclusive events – first friend hit the target and the second has not; and second friend hit the target and first has not. Since the events of hitting are independent (note that independence is preserved when taking complements) is the probability given by the product of the probabilities of given elementary elements. That is,

$$P = 0.4 \cdot (1 - 0.3) + (1 - 0.4) \cdot 0.3 = 0.46.$$ □

1.G.40. We flip three coins twelve times. What is the probability that at least one flipping results in three tails?

Solution. If we realize that when repeating the flipping, the individual results are independent, and denote for $i \in \{1, \dots, 12\}$ by A_i the event „the i-th flipping results in three tails", we are determining

$$P\left(\bigcup_{i=1}^{12} A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_{12})).$$

For every $i \in \{1, \dots, 12\}$ is $P(A_i) = 1/8$, since at every coin of the three the tail is with the probability $1/2$ independently of the results of the other coins. Now we can write the final probability

$$1 - \left(\tfrac{7}{8}\right)^{12}.$$ □

1.G.41. 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? (The coin is "fair".)

Solution. There are 2^{200} of possible results of the elections (considered to be sequences 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

$$\frac{\binom{200}{100}}{2^{200}} = \frac{\frac{200!}{100! \cdot 100!}}{2^{200}} \doteq 0.056.$$

□

1.G.42. Seven Czechs and five English are randomly divided into two (nonempty) groups. What is the probability that one group consists of Czechs only?

Solution. There are $2^{12} - 1$ of possible divisions. If one group consists of Czechs only, it means that all English are in one group (either in the first or in the second). It remains to divide the Czechs into two nonempty groups, that can be done in $2^7 - 1$ ways. In the end we must add 1 for the division which puts all English in one group and all Czechs in another,

$$\frac{2 \cdot (2^7 - 1) + 1}{2^{12} - 1}$$

□

1.G.43. 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 event „at i-th drawing the ace was drawn". Since the individual events A_i are (stochastically) independent, we know that

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n))$$

for every $n \in \mathbb{N}$. We are looking for an $n \in \mathbb{N}$ such that it holds that

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n)) > 0.9.$$

Clearly is $P(A_i) = 1/10$ for any $i \in \mathbb{N}$. Thus it is enough to solve the equation

$$1 - \left(\tfrac{9}{10}\right)^n > 0.9,$$

from which we can express

$$n > \frac{\log_a 0.1}{\log_a 0.9}, \quad \text{kde } a > 1.$$

Evaluating, we obtain that we must do the drawing at least twenty two times. □

1.G.44. Texas hold'em. Let us now solve a couple of simple exercises concerning the popular card game Texas hold'em, whose rules we will not state (if the reader does not know them, she can look them up on the Internet). What is the probability that

  i) the starting combination is a tuple of the same symbols?
 ii) in my starting tuple of cards there is an ace?
iii) in the end I have one of the six best combinations of cards?
iv) I win, if I hold in my hand ace and a triple of twos (of any colour), on the flop there is ace and two twos and on the turn there is a third three and all these four cards have distinct colour? (The last card river is not yet turned)

Solution.

  i) The number of distinct symbols is 13 and there are always four of them (one of each colour). Thus the number of tuples with the same symbols is $13\binom{4}{2} = 78$. The number of all possible tuples is $\binom{13 \cdot 4}{2} = 1326$. The probability of having same symbols is then $\frac{1}{17} \doteq 0.06$.
 ii) One card is the ace, that is four choices, and the second is arbitrary, that is 51 choices. But we have counted twice the tuples with two aces, of which there are $\binom{4}{2} = 6$. Thus we obtain $4 \cdot 51 - 6 = 198$ tuples and the probability is $\frac{198}{1326} \doteq 0.15$.

55

iii) Let us compute the probabilities of the individual best combinations:

ROYAL FLUSH: There are exactly only four such combinations – one of each colours. The number of combinations of five cards are $\binom{52}{5} = 2598960$. The probability is thus equal to $1.5 \cdot 10^{-6}$. Very small :)

STRAIGHT FLUSH: Sequence which ends with the highest card in the range 6 to K, that is eight choices for every colours. We obtain $\frac{32}{2598960} \doteq 1.2 \cdot 10^{-5}$.

POKER: Four identical symbols – 13 choices (for every symbol one). The fifth card can be arbitrary, that is 48 choices. That makes $\frac{624}{2598960} \doteq 2.4 \cdot 10^{-4}$.

FULL HOUSE: Three identical symbols make $13\binom{4}{3} = 52$ choices and two identical symbols make $12\binom{4}{2} = 72$ choices. The probability is $\frac{3744}{2598960} \doteq 1.4 \cdot 10^{-3}$.

FLUSH: All five cards of the same colour means $4\binom{13}{5} = 5148$ choices and the probability is then $\frac{5148}{2598960} \doteq 2 \cdot 10^{-3}$.

STRAIGHT: The highest card of the sequence is in the range from 6 to Ace, that is 9 choices. The colour of every card is arbitrary, that makes $9 \cdot 4^5 = 9216$ choices. But we have counted both straight flush and royal flush which we must subtract.

For determining the probability of one of the six best combinations we don't have to do that, we just do not count the first two combinations. Therefore we obtain the probability approximately $3.5 \cdot 10^{-3} + 2 \cdot 10^{-3} + 1.4 \cdot 10^{-3} + 2.4 \cdot 10^{-4} = 7.14 \cdot 10^{-3}$.

iv) The situation is clearly pretty good and therefore it will be better to count bad situation, that is, when the opponent has even better combination. I have at this moment full house of two aces and three two's. The only combination that could beat me at this moment is either full house of three aces and two twos or a poker of twos. That means that the enemy must have either the ace or the last two. If he has the two and any other card, then he clearly wins no matter what card is river. How many ways are there for this other card in his hand? $3 + 4 + \cdots + 4 + 2 = 45$ (one triple and two aces cannot be in his hand since I have them). There are $\binom{46}{2} = 1035$ remaining combination and the probability of such loss is then 0.043. If he has an ace in his hand, then the following can happen. If he holds two aces, then he again wins if two is not on the river – then I would have split poker. The probability of my (conditional) loss is then $\frac{1}{1035} \cdot \frac{43}{44} \doteq 10^{-3}$. If the enemy has in his hand ace and some other card than 2 and A, then it is a draw no matter what is on the river. The total probability of the win is thus almost 96 %. $\qquad \square$

1.G.45. A volleyball team (with libero, that is, 7 people) sits after a match in a pub and drinks beer. But there is not enough mugs, and thus the publican keeps using the same seven. What is the probability that

i) exactly one person does not receive the mug he had last round,

ii) nobody receives the mug he had last round,

iii) exactly three receive the mug they had last round.

Solution.

i) If six people receive the mug they had last round, then clearly the seventh person also receives the mug he had last round, the probability is thus zero.

ii) Let M is the set of all orderings and event A_i occurs when the i-th person receives his mug from last round. We want to calculate $|M - \cup_i A_i|$. We obtain $7! \sum_{k=0}^{7} \frac{(-1)^k}{k!} = 1854$. And the probability is $\frac{1854}{5040} = \frac{103}{280} \doteq 0.37$.

iii) We choose which three receive the mug they had last round – $\binom{7}{3} = 35$ choices. The remaining four must receive mugs from somebody else. That is again the formula from the previous section, specifically it is $4! \sum_{k=0}^{4} \frac{(-1)^k}{k!} = 9$ choices. In total we have $9 \cdot 35 = 315$ choices and the probability is $\frac{315}{5040} = \frac{1}{16}$. $\qquad \square$

1.G.46. In how many ways can we place n identical rooks on a chessboard $n \times n$ such that every non-occupied position is threatened by some of the rooks?

Solution. Such placements are a union of two sets: the set of placements where in at least one row there is one rook (therefore in every row there is exactly one; this set has n^n elements – in every row we choose independently one position for the rook), and the set of placements where in every column there is at least one (that is exactly one) rook (as before, this set has n^n elements). The intersection of these sets has $n!$ elements (the places for the rooks are chosen sequentially starting in the first row – there we have n choices, in the second only $n-1$ – one column is already occupied...). Using the inclusion-exclusion principle, we obtain

$$2n^n - n!.$$

$\square$

1.G.47. Determine the probability that when throwing two dice at least one resulted in four, if the sum is 7.

Solution. We solve this exercise using the classical probability, where the condition is interpreted as restriction of the probability space. The space has due to the condition 6 elements, and exactly 2 of those are favourable to the given event. The answer is thus $2/6 = 1/3$.

$\square$

1.G.48. We throw two 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.

Solution. 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.

$\square$

1.G.49. Let us have 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 (which have the same probability), and exactly four of them are favourable. Thus we see that the probability is $1/8$. In the second case when we do not return the card, is probability also the same. It is enough to consider that when drawing all the cards one by one is the probability of the ace as the first card identical to the probability that the ace is the second card. We could also use conditional probability, that results into

$$\frac{4}{32} \cdot \frac{3}{31} + \frac{28}{32} \cdot \frac{4}{31} = \frac{1}{8}.$$

$\square$

1.G.50. Consider families with two children and for simplicity assume that all choices in the set $\Omega = \{bb, bg, gb, gg\}$, where b stands for „boy" and g stands for „girl" (considering the age of the children) have the same probability. Choose random events

$$H_1 \text{ – family has a boy,} \quad A_1 \text{ – family has two boys.}$$

Compute $P(A_1|H_1)$.

Similarly consider families with three children, where

$$\Omega = \{bbb, bbg, bgb, gbb, bgg, gbg, ggb, ggg\}.$$

If

$$H_2 \text{ – the family has both boy and girl,} \quad A_2 \text{ – the family has at most one girl,}$$

decide whether the events A_2 and H_2 are independent.

Solution. Considering which of the four elements of the set Ω are (not) favourable to the event A_1 or H_1, we easily obtain

$$P(A_1|H_1) = \frac{P(A_1 \cap H_1)}{P(H_1)} = \frac{P(A_1)}{P(H_1)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Further we have to determine whether the following holds:

$$P(A_2 \cap H_2) = P(A_2) \cdot P(H_2).$$

Again we just have to realize that exactly the elements *bbb*, *bbg*, *bgb*, *gbb* of the set Ω, are favourable to the event A_2; to the event H_2 the elements *bbg*, *bgb*, *gbb*, *bgg*, *gbg*, *ggb* are favourable and to the event $A_2 \cap H_2$ the elements *bbg*, *bgb*, *gbb*. Therefore

$$P(A_2 \cap H_2) = \tfrac{3}{8} = \tfrac{4}{8} \cdot \tfrac{6}{8} = P(A_2) \cdot P(H_2),$$

which means that the events A_2 and H_2 are independent. □

1.G.51. 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. Express the probability that in the hat there is more black balls than white balls, if there is at least one black ball in the hat.

Solution. Let us have the following two events

A – there are more black balls than white balls in the hat,

H – there is at least one black ball in the hat.

We want to express $P(A|H)$. Note that the probability $P(H^C)$ of the complementary event to the event H is 2^{-5} and that the probability of the event is the same as the probability $P(A^C)$ of the complementary event (there are more white balls in the hat). Necessarily, $P(H) = 1 - 2^{-5}$, $P(A) = 1/2$. Furthermore $P(A \cap H) = P(A)$, since the event H contains the event A (the event A has H as a consequence). Thus we have obtained

$$P(A|H) = \tfrac{P(A \cap H)}{P(H)} = \tfrac{\frac{1}{2}}{1-(\frac{1}{2})^5} = \tfrac{16}{31}.$$ □

1.G.52. 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.

Solution. We solve this exercise using the theorem about multiplication of probabilities. 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

$$\tfrac{9}{16} \cdot \tfrac{8}{15} \cdot \tfrac{7}{14} = 0.15.$$ □

1.G.53. In the box there are 10 balls, 5 of them are black and 5 are white. We will sequentially draw the balls, and we do not return them back. Determine the probability that first we draw a white ball, then a black, then a white and in the last, fourth turn again a white.

Solution. We use the theorem about multiplication of probabilities. In the first round we draw a white ball with the probability $5/10$, then a black ball with probability $5/9$, then a white ball with probability $4/8$ and in the end a white ball with probability $3/7$. That gives

$$\tfrac{5}{10} \cdot \tfrac{5}{9} \cdot \tfrac{4}{8} \cdot \tfrac{3}{7} = \tfrac{5}{84}.$$ □

1.G.54. From a deck of 32 cards we randomly draw six cards. Compute the probability that the first king will be chosen as the sixth card (that is, the previous five cards do not contain any king).

Solution. Using the theorem about multiplication of probabilities we have

$$\tfrac{28}{32} \cdot \tfrac{27}{31} \cdot \tfrac{26}{30} \cdot \tfrac{25}{29} \cdot \tfrac{24}{28} \cdot \tfrac{4}{27} \doteq 0.0723.$$ □

1.G.55. What is the probability that a sum of two randomly chosen positive numbers smaller than 1 is smaller than 3/7?

Solution. It is clear that it is a simple exercise on geometrical probability where the basic space Ω is a square with vertices at $[0, 0]$, $[1, 0]$, $[1, 1]$, $[0, 1]$ (we are choosing two numbers in $[0, 1]$). We are interested in the probability of the event that a randomly chosen point $[x, y]$ in this square satisfies $x + y < 3/7$, that is, the probability that the point lies in the triangle A with vertices at $[0, 0]$, $[3/7, 0]$, $[0, 3/7]$. Now we can easily compute

$$P(A) = \tfrac{\text{vol } A}{\text{vol } \Omega} = \tfrac{(\frac{3}{7})^2 / 2}{1} = \tfrac{9}{98}.$$

$\square$

1.G.56. Let a pole be randomly broken into three parts. Determine the probability that the length of the second (middle) part is greater than two thirds of the length of the pole before the breaking.

Solution. Let d stand for the length of the pole. The breaking of the pole at two points is given by the choice of the points where we split the pole. Let x be the point which is the first (closer to left end of the pole), and $x + y$ be the point where the second splitting occurs. That says that the basic space is the set $\{[x, y]; x \in (0, d), y \in (0, d - x)\}$, that is, a triangle with vertices at $[0, 0]$, $[d, 0]$, $[0, d]$. The length of the middle part is given by the value of y. The condition from the exercise statement can be now restated as $y > 2d/3$, which corresponds to the triangle with vertices at $[0, 2d/3]$, $[d/3, 2d/3]$, $[0, d]$. Areas of the considered triangles are $d^2/2$ a $(d/3)^2/2$, therefore the probability is

$$\frac{\frac{d^2}{3^2 \cdot 2}}{\frac{d^2}{2}} = \tfrac{1}{9}.$$

$\square$

1.G.57. A pole of length 2 m is randomly divided into three parts. Determine the probability of the event that the third part is shorter than $1, 5$ m.

Solution. This exercise is for using the geometrical probability, where we are looking for the probability that the sum of the lengths of the first two parts is greater than one fourth of the length of the pole. We determine the probability of the complementary event, that is, the probability that if we randomly choose two points on the pole, both of them are in the first quarter of the pole. The probability of this event is $1/4^2$, since the probability of picking a point in the first quarter of the pole is clearly $1/4$ and this choice is independently repeated (once). Thus the probability of the complementary event is $15/16$. $\square$

1.G.58. Mirek and Marek have a lunch at the school canteen. The canteens opens from 11 to 14. Each of them eats the 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?

Solution. The space of all possible events is a square 3×3. Denote by x the arrival time of Mirek and by y the arrival time of Marek, these two meet if and only if $|x - y| \le 1/2$. This inequality determines in the square of possible events the area whose volume is $11/36$ of the volume of the whole square. Thus that is also the probability of the event. $\square$

1.G.59. From Brno Honza rides a car to Prague randomly between 12 and 16, and in the same time interval Martin rides a car to Brno from Prague. Both stop in a motorest in the middle of the trip for thirty minutes. What is the probability that they meet there, if Honza's speed is 150 km/h and Martin's is 100 km/h? (The distance Praha-Brno is 200 km).

Solution. If we denote the departure time of Martin by x and the departure time of Honza by y, and in order to have fewer fractions in the following calculations choose a time unit to be ten minutes, then the base space is a square 24×24. The arrival time of Martin to the motorest is $x + 6$, arrival time of Honza is $y + 4$. As in the previous exercise, the event that they meet in the motorest is equivalent to the event that their arrival times do not differ by more than thirty minutes, that is, $|(x + 6) - (y + 4)| \le 3$. This condition determines an area with volume $24^2 - \tfrac{1}{2}(23^2 + 19^2)$ (see the figure)
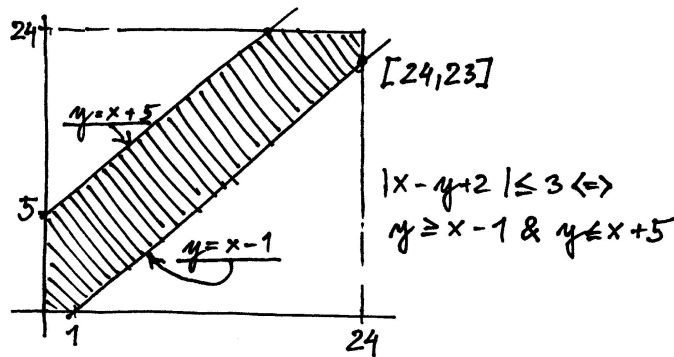
and the probability

$$p = \frac{24^2 - \frac{1}{2}(23^2 + 19^2)}{24^2} = \frac{131}{576} \doteq 0.227.$$

$\square$

1.G.60. Mirek departs randomly between 10 and 20 o'clock from Brno to Prague. Marek departs randomly in the same interval from Prague to Brno. The trip takes 2 hours. What is the probability that they meet on the road (they use the same road)?

Solution. We are solving analogously to the previous exercise. The space of all events is a square 10×10, Mirek, departing at the time x, meets Marek, departing at the time y if and only if $|x - y| \leq 2$. The probability is $p = \frac{36}{100} = \frac{9}{25} = 0.36$.

$\square$
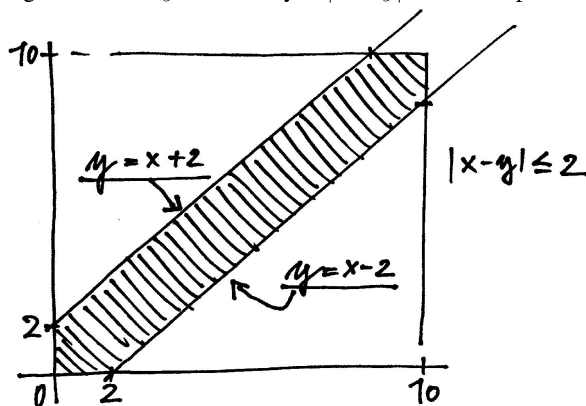
1.G.61. Two meter-long pole is randomly divided into three pieces. Determine the probability that a triangle can be built of the pieces.

Solution. Division of the pole is given as in the previous exercises by the points of cutting x and y and the probability space is again a square 2×2. In order to be able to build a triangle of the pieces, the lengths of the parts must satisfy the triangle inequalities, that is, sum of lengths of any two parts must be greater than the length of the third part. Since the sum of the lengths is 2 meters, this condition is equivalent to the condition that each part must be smaller than 1 meter. Using the cut-points x and y, we can express this that it cannot simultaneously hold $x \leq 1$ and $y \leq 1$ or simultaneously $x \geq 1$ and $y \geq 1$ (this corresponds to the conditions that the border parts of the pole are smaller than 1), and also $|x - y| \leq 1$ (the middle part is smaller than one). These conditions are satisfied by the shaded area in the picture, whose volume is $1/4$.
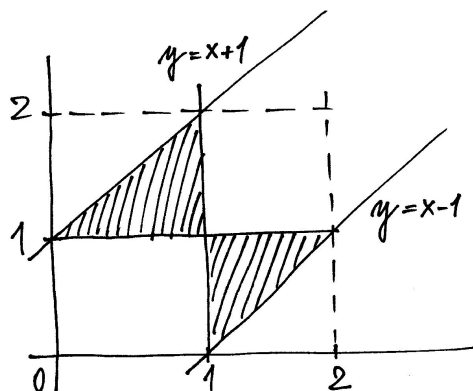
1.G.62. Do the equations

(a) $\begin{aligned} 4x_1 &- \sqrt{3}x_2 &= 3, \\ x_1 &- 2\sqrt{7}x_2 &= -2; \end{aligned}$

(b) $\begin{aligned} 4x_1 &- \sqrt{3}x_2 &= 16, \\ x_1 &- 2\sqrt{7}x_2 &= -7; \end{aligned}$

(c) $\begin{aligned} 4x_1 &+ 2x_2 &= 7, \\ -2x_1 &- x_2 &= -3 \end{aligned}$

have a unique solution (that is, exactly one)?

Solution. The set of equation is uniquely solvable if and only if the determinant of the matrix given by the left-hand side coefficients is nonzero. Therefore, the coefficients on the right-hand side do not influence the uniqueness of the solution. Thus we have to have the same answer in (a) and (b). Since

$$\begin{vmatrix} 4 & -\sqrt{3} \\ 1 & -2\sqrt{7} \end{vmatrix} = 4 \cdot (-2\sqrt{7}) - (-\sqrt{3} \cdot 1) \neq 0,$$

$$\begin{vmatrix} 4 & 2 \\ -2 & -1 \end{vmatrix} = 4 \cdot (-1) - (2 \cdot (-2)) = 0,$$

for (a) and (b) there is a unique solution and in (c) there is not. If we multiply the second equation in (c) by -2, we see that it has no solution at all. $\square$

1.G.63. Determine $A \cdot A$ for

$$A = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}, \quad \text{where } \varphi \in \mathbb{R}.$$

Solution. We know that the mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}, \quad x, y \in \mathbb{R}$$

is the rotation of the plane $\mathbb{R}^2$ around the origin through the angle φ in the positive direction. Since matrix multiplication is associative, we obtain that the mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \cdot \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}, \quad x, y \in \mathbb{R}$$

is a rotation through the angle 2φ. That means that

$$A \cdot A = \begin{pmatrix} \cos 2\varphi & -\sin 2\varphi \\ \sin 2\varphi & \cos 2\varphi \end{pmatrix}.$$

Note that we could have directly multiplied $A \cdot A$ (and apply the formulas for sine and cosine of double angle). But repeating the aforementioned method (or using the mathematical induction) yields

$$A^n = \begin{pmatrix} \cos n\varphi & -\sin n\varphi \\ \sin n\varphi & \cos n\varphi \end{pmatrix}, \quad n = 2, 3, \ldots,$$

easier (we set $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$, etc.). $\square$

1.G.64. **The parallelogram identity.** The calculation in coordinates can be useful in plane geometry. Let us demonstrate this on the proof "parallelogram identity": if $u, v \in \mathbb{R}^2$, then:

$$2(\|u\|^2 + \|v\|^2) = \|u + v\|^2 + \|u - v\|^2.$$

Thus the sum of the squares of the diagonals of a parallelogram is the sum of the squares of the lengths of the four sides of the parallelogram.

Solution. Writing both sides of the equation into the coordinates $u = (u_1, u_2)$, $v = (v_1, v_2)$ yields:

$$\|u + v\|^2 + \|u - v\|^2$$
$$= (u_1 + v_1)^2 + (u_2 + v_2)^2 + (u_1 - v_1)^2 + (u_2 - v_2)^2$$
$$= u_1^2 + 2u_1v_1 + v_1^2 + u_2^2 + 2u_2v_2 + v_2^2 +$$
$$\quad + u_1^2 - 2u_1v_1 + v_1^2 + u_2^2 - 2u_2v_2 + v_2^2$$
$$= 2(u_1^2 + u_2^2 + v_1^2 + v_2^2)$$
$$= 2(\|u\|^2 + \|v\|^2).$$

$\square$

1.G.65. Compute the area S of a quadrilateral given by the vertices

$$[0, -2], \quad [-1, 1], \quad [1, 5], \quad [1, -1].$$

Solution. In the usual notation

$$A = [0, -2], \quad B = [1, -1], \quad C = [1, 5], \quad D = [-1, 1]$$

and the usual division of the quadrilateral into triangles ABC and ACD with areas S_1 and S_2 we obtain

$$S = S_1 + S_2 = \tfrac{1}{2} \begin{vmatrix} 1 - 0 & 1 - 0 \\ -1 + 2 & 5 + 2 \end{vmatrix} + \tfrac{1}{2} \begin{vmatrix} 1 - 0 & -1 - 0 \\ 5 + 2 & 1 + 2 \end{vmatrix} = \tfrac{1}{2}(7 - 1) + \tfrac{1}{2}(3 + 7) = 8.$$

$\square$

1.G.66. Determine the area of the quadrilateral $ABCD$ with vertices $A = [1, 0]$, $B = [11, 13]$, $C = [2, 5]$ a $D = [-2, -5]$.

Solution. We divide the quadrilateral into two triangles ABC and ACD. We compute their areas by computing absolute values of the determinants, see 1.5.11,

$$S = \left| \frac{1}{2} \cdot \begin{vmatrix} 1 & 5 \\ 10 & 13 \end{vmatrix} \right| + \left| \frac{1}{2} \cdot \begin{vmatrix} 1 & 5 \\ -3 & -5 \end{vmatrix} \right| = \frac{47}{2}.$$

$\square$

1.G.67. 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

$$\begin{vmatrix} 6 - 5 & 6 - 5 \\ 8 - 5 & 9 - 5 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 3 & 4 \end{vmatrix} = 1 \cdot 4 - 1 \cdot 3 = 1.$$

$\square$

1.G.68. Give the area of a meadow, which is determined on the area map by the points at positions $[-7, 1]$, $[-1, 0]$, $[29, 0]$, $[25, 1]$, $[24, 2]$ and $[17, 5]$. (Ignore the measurement units. They are determined by the ratio of the area map to the reality.)

Solution. The given hexagon can be divided into four triangles with vertices at

$$[-7, 1], [-1, 0], [17, 5]; \qquad [-1, 0], [24, 2], [17, 5];$$
$$[-1, 0], [25, 1], [24, 2]; \qquad [-1, 0], [29, 0], [25, 1].$$

The areas are 24, 89/2, 27/2 and 15 respectively, which gives the total area as

$$24 + 44\tfrac{1}{2} + 13\tfrac{1}{2} + 15 = 97.$$

$\square$

1.G.69. Determine the area of a triangle $A_2A_3A_{11}$, where $A_0A_1 \dots A_{11}$ are the vertices of a regular dodecagon inscribed in a circle of radius 1.

Solution. The vertices of the dodecagon can be identified with the twelfth roots of 1 in the complex plane. As in **??** we find out $A_2 = \cos(\pi/3) + i\sin(\pi/3) = 1/2 + i\sqrt{3}/2$, $A_3 = \cos(\pi/2) + i\sin(\pi/2) = i$, $A_{11} = \cos(-\pi/6) + i\sin(-\pi/6) = \sqrt{3}/2 - i/2$, that means the that the coordinates of these points in the complex plane are $A_2 = [1/2, \sqrt{3}/2]$, $A_3 = [0, 1]$, $A_{11} = [\sqrt{3}/2, -\frac{1}{2}]$. According to the formula for the area of a triangle, the area of the triangle S is

$$S = \frac{1}{2}\begin{vmatrix} A_2 - A_{11} \\ A_3 - A_{11} \end{vmatrix} = \frac{1}{2}\begin{vmatrix} \frac{1}{2} - \frac{\sqrt{3}}{2} & \frac{1}{2} + \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{3}{2} \end{vmatrix} = \frac{3 - \sqrt{3}}{4}.$$

$\square$

1.G.70. Determine which sides of the quadrilateral with vertices $A = [95, 99]$, $B = [130, 106]$, $C = [40, 60]$, $D = [130, 120]$, are visible from the point $[2, 0]$. $\bigcirc$

1.G.71. 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, $1 + 4 \cdot 1 + \binom{4}{2} \cdot 2 + \binom{4}{3} \cdot 5 + 15 = 52$. $\square$

1.G.72. Determine the number of ordering relations over a three-element set. $\bigcirc$

1.G.73. Determine the numer of ordering relations over the set $\{1, 2, 3, 4\}$ such that the elements 1 and 2 are not comparable (that is, neither $1 \prec 2$ nor $2 \prec 1$, where $\prec$ stands for the ordering relation). $\bigcirc$

1.G.74. Determine 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, as we know from the previous exercise. $\square$

1.G.75. Give all the elements in $S \circ R$, if

$$R = \{(2, 4), (4, 4), (4, 5)\} \subset \mathbb{N} \times \mathbb{N},$$
$$S = \{(3, 1), (3, 2), (3, 5), (4, 1), (4, 4)\} \subset \mathbb{N} \times \mathbb{N}.$$

Solution. Considering all choices of two ordered tuple

$$(2, 4), (4, 1); \quad (2, 4), (4, 4); \quad (4, 4), (4, 1); \quad (4, 4), (4, 4)$$

satisfying that the second element of the first ordered tuple—which is a member of R—equals the first element of the second ordered tuple—which is a member of S—we obtain

$$S \circ R = \{(2, 1), (2, 4), (4, 1), (4, 4)\}.$$

$\square$

1.G.76. Let a binary relation be given

$$R = \{(0, 4), (-3, 0), (5, \pi), (5, 2), (0, 2)\}$$

between sets $A = \mathbb{Z}$ a $B = \mathbb{R}$. Express R^{-1} and $R \circ R^{-1}$.

Solution. We can immediately see that

$$R^{-1} = \{(4, 0), (0, -3), (\pi, 5), (2, 5), (2, 0)\}.$$

Furthermore,

$$R \circ R^{-1} = \{(4, 4), (0, 0), (\pi, \pi), (2, 2), (4, 2), (\pi, 2), (2, \pi), (2, 4)\}.$$

$\square$

1.G.77. Decide whether the relation R determined by the condition:

(a) $(a, b) \in R \iff |a| < |b|$;

(b) $(a, b) \in R \iff |a| = |2b|$

over the set of integers $\mathbb{Z}$ is transitive.

Solution. In the first case R is transitive, because

$$|a| < |b|, |b| < |c| \implies |a| < |c|.$$

In the second case R is not transitive. For instance, consider

$$(4, 2), (2, 1) \in R, \quad (4, 1) \notin R.$$

$\square$

1.G.78. Find all relations over $M = \{1, 2\}$, which are not antisymmetric. Which of them are transitive?

Solution. There are four relations that are not antisymmetric. They are exactly subsets of the set $\{1, 2\} \times \{1, 2\}$, which contain the elements $(1, 2), (2, 1)$ (otherwise the condition of antisymmetry is satisfied). Of these four only the relation

$$\{(1, 1), (1, 2), (2, 1), (2, 2)\} = M \times M,$$

is transitive, because not containing tuples $(1, 1)$ and $(2, 2)$ in a transitive relation means that the relation cannot contain both $(1, 2)$ and $(2, 1)$.

$\square$

1.G.79. We have a set $\{3, 4, 5, 6, 7\}$. Write explicitly the relations

 i) a divides b,

 ii) Either a divides b or b divides a,

 iii) a and b have a common divisor greater than one,

and examine their properties.

$\bigcirc$

1.G.80. Is there an equivalence relation, which is also an ordering, over the set of all lines in the plane?

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). That means that a line can be in relation only with itself. If we define the relation such that two lines are in relation if and only if they are identical, we obtain "very natural" relation which is both equivalence relation and ordering. We just need to check that it is transitive, which it trivially is. Thus the only relation satisfying the problem statement is the identity over the set of all lines in the plane.

$\square$

1.G.81. Determine whether the relation

$$R = \{(k, l) \in \mathbb{Z} \times \mathbb{Z}; \, |k| \geq |l|\}$$

over the set $\mathbb{Z}$ is an equivalence and/or an ordering.

Solution. The relation R is not an equivalence: it is not symmetric (take $(6, 2) \in R$, $(2, 6) \notin R$); it is not an ordering: it is not antisymmetric (take $(2, -2) \in R$, $(-2, 2) \in R$).

$\square$

1.G.82. Show that the intersection of any equivalence relation over a set X is again an equivalence relation, and that the union of two ordering relations over a set X does not have to be an ordering.

Solution. We see that the intersection of equivalence relations is reflexive, symmetric and transitive: all the equivalence relations must contain the tuple (x, x) for every $x \in X$, therefore the intersection contains that tuple too. If the element (x, y) is in the intersection, then the element (y, x) is also in the intersection (just use the fact that every equivalence is symmetric). If tuples (x, y) and (y, z) are in the intersection, then both are in the equivalences also. Since the equivalences are transitive, they all contain the element (x, z) and thus that element is also in the intersection.

   If we chose $X = \{1, 2\}$ and the ordering relation

$$R_1 = \{(1, 1), (2, 2), (1, 2)\}, \quad R_2 = \{(1, 1), (2, 2), (2, 1)\}$$

over X, we obtain the relation

$$R_1 \cup R_2 = \{(1,1), (2,2), (1,2), (2,1)\},$$

which is not antisymmetric, thus not an ordering. □

1.G.83. Over the set $M = \{1, 2, \ldots, 19, 20\}$ there is an equivalence relation $\sim$ such that $a \sim b$ for any $a, b \in M$ if and only if the first digits of the numbers a, b are the same. Construct the partition given by this equivalence.

Solution. Two numbers from the set M are in the same equivalence class if and only if they are in the relation (first digit is the same). Therefore the partition consists of the sets

$$\{1, 10, 11, \ldots, 18, 19\}, \{2, 20\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}.$$
□

1.G.84. We are given partition of two classes $\{b, c\}$, $\{a, d, e\}$ of the set $X = \{a, b, c, d, e\}$. Write down the equivalence relation R over the set X which gives this partition.

Solution. Equivalence R is determined by the fact that the two elements are in relation if and only if they are in the same partition class (note also that R must be symmetric), and every element is in relation with itself (R must be reflexive). Therefore R contains exactly

$$(a, a), (b, b), (c, c), (d, d), (e, e),$$
$$(b, c), (c, b), (a, d), (a, e), (d, a), (d, e), (e, a), (e, d).$$
□

1.G.85. 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). □

1.G.86. What is the maximal domain $D \subseteq \mathbb{R}$ and codomain $H \subseteq \mathbb{R}$ such that the following mappings are bijective, and what is then the inverse function?

i) $x \mapsto x^4$

ii) $x \mapsto x^3$

iii) $x \mapsto \frac{1}{x+1}$

Solution.

i) $D = [0, \infty)$ and $H = [0, \infty)$ or also $D = (-\infty, 0]$ a $H = [0, \infty)$. The inverse function is then $x \mapsto \sqrt[4]{x}$.

ii) $D = H = \mathbb{R}$ and the inverse function is $x \mapsto \sqrt[3]{x}$.

iii) $D = \mathbb{R} \smallsetminus \{-1\}$ and $H = \mathbb{R} \smallsetminus \{0\}$. The inverse function is $x \mapsto \frac{1}{x} - 1$. □

1.G.87. Consider a relation $\mathbb{R} \times \mathbb{R}$. A point is in the relation whenever it holds that

$$(x - 1)^2 + (y + 1)^2 = 1.$$

Can we describe the points using the function $y = f(x)$? Depict the points in the relation.

Solution. We cannot, because for instance $y = -1$ has two preimages: $x = 0$ and $x = 2$. The points lie on a circle with the centre at the point $(1, -1)$ and radius 1. □

1.G.88. Let for any two integers k, l hold that $(k, l) \in R$ whenever the number $4k - 4l$ is an integral multiple of 7. Is such a relation over R an equivalence? Is it an ordering?

Solution. Note that two integers are in the relation R if and only if they have the same remainder under the division by 7. Therefore it is an example of the so-called remainder class of integers. Therefore we know that the relation R is an equivalence relation. Its symmetry (for instance, $(3, 10), (10, 3) \in R, 3 \neq 10$) implies that it is not an ordering. □

1.G.89. Let a relation R be defined over the set $N = \{3, 4, 5, \ldots, n, n+1, \ldots\}$, such that two numbers are in the relation whenever they are relatively prime (that is, the prime decompositions of the numbers do not contain any common number). Determine whether this relation is reflexive, symmetric, antisymmetric, transitive.

Solution. For a tuple of the same numbers it holds that $(n, n) \notin R$. Therefore the relation is not reflexive. It is clear that when two numbers are relatively prime or not, it does not matter how they are ordered – it is a property of unordered tuples. Therefore, R is symmetric. From the symmetry we have that it is not antisymmetric (for instance, $(3, 5) \in R, 3 \neq 5$). Since R is symmetric and $(n, n) \notin R$ for any number $n \in N$, a choice of two distinct numbers which are in the relation gives that R is not transitive. □

1.G.90. Determine the number of injective mappings of the set $\{1, 2, 3\}$ to the set $\{1, 2, 3, 4\}$.

Solution. Any injective mapping among the given sets is given by choosing an (ordered) triple from the set $\{1, 2, 3, 4\}$ (the elements in the chosen triple will correspond in order to images of the numbers $1, 2, 3$) and vice versa. Every injective mapping gives such a triple. Thus the number of injective mappings equals the number of ordered triples among four elements, that is $v(3, 4) = 4 \cdot 3 \cdot 2 = 24$. □

1.G.91. How many relations are there over an n-element set?

Solution. A relation is an arbitrary subset of the cartesian product of the set with itself. This cartesian product has n^2 elements, thus the number of all relations over an n-element set is 2^{n^2}. □

1.G.92. How many reflexive relations are there over an n-element set?

Solution. The relation over the set M is reflexive if and only if it has the diagonal relation $\Delta_M = \{(a, a), \text{ all } a \in M\}$ as a subset. As for the rest of the $n^2 - n$ ordered pairs in the cartesian product $M \times M$, we have independent choice, whether or not the pair belongs to the relation. In total we have 2^{n^2-n} different reflexive relations over an n-element set. □

1.G.93. How many symmetric relations are there over an n-element set?

Solution. A relation R over the set M is symmetric if and only if the intersection of R with each $\{(a, b), (b, a), \text{ where } a \neq b, a, b \in M\}$ is either the whole two-element set or is empty. There are $\binom{n}{2}$ two-element subsets of the set M. If we also declare what the intersection of R and the diagonal relation $\Delta_M = \{(a, a), \text{ where } a \in M\}$ should be, then R is completely determined. In total we are to do $\binom{n}{2} + n$ independent choices between two alternatives: each set of the type $\{(a, b), (b, a)| \text{ where } a, b \in M, a \neq b\}$ is either the subset of R or it is disjoint with R. Every pair $(a, a), a \in M$ is either in R or not. In total we have $2^{\binom{n}{2}+n}$ symmetric relations over an n-element set. □

1.G.94. How many anti-symmetric relations over an n-element set are there?

Solution. A relation R over the set M is anti-symmetric if and only if the intersection of R with each set $\{(a, b), (b, a)\}, a \neq b, a, b \in M$ is either empty or one-element (which means that it is either $\{(a, b)\}$ or $\{(b, a)\}$ but not both). The intersection of R with the diagonal relation is arbitrary. By declaring what these intersections are, the relation R is completely determined. In total we have $3^{\binom{n}{2}}2^n$ anti-symmetric relations over an n-element set. □

1.G.95. Determine the number of ordering relations of the set $\{1, 2, 3, 4, 5\}$ such that exactly two pairs of element are incomparable. ○

Solution to the exercises

1.B.3. $y_n = 2(\frac{3}{2})^n - 2$.

1.G.21.

  i) $2^6 = 64$

  ii) $\binom{6}{4} = 15$

  iii) 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. 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.70. First, we orient the vertices of the given quadrangle in the counter-clockwise order: $ABCD$. After computing the corresponding determinants as in the previous exercises we see that only the side CB is visible.

1.G.72. 19.

1.G.73. 87.

1.G.79.

  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 $i = 1, \ldots, 7$ and additionally $(3,6), (6,3)$, check that it is an equivalence relation.

  iii) (i, i) for $i = 1, \ldots, 7$ and also $(3,6), (6,3), (4,6), (6,4)$. Check that it is not an equivalence, since transitivity does not hold.

1.G.95. Three different Hasse diagrams which satisfy the given condition. In total $5! + 5! + 5!/4 = 270$.

CHAPTER 2

Elementary linear algebra

Can't you count with scalars yet?

– no worry, let us go straight to matrices...
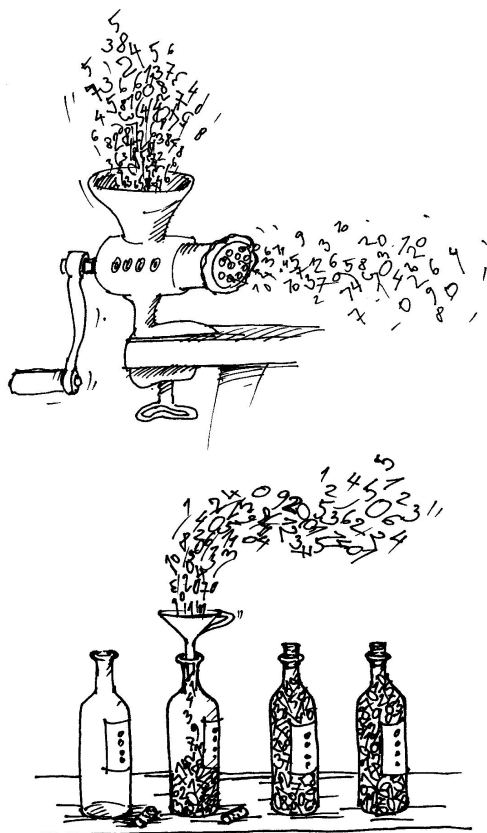
A. Vectors and matrices

Before we approach vector spaces in a systematic way, we begin with something we know – systems of linear equations. Step by step we shall see that the vector spaces are hidden behind them and we get used to the matrix calculus on the way.

2.A.1. 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

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 tools.

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 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 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.

by mixing the colours he usually sells (he has enough in his warehouse). He has

- 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. Denote by

- x – the number of litres of reddish colour to be used;
- y – the number of litres of bluish colour to be used;
- z – the number of litres) greenish colour to be used;

By mixing the colours we want a colour that contains 270 litres of red. Note that reddish contains 50 % red, greenish contains 12,5 % red and bluish 20 % red. Thus the following has to be satisfied:

$$0,5x \quad + \quad 0,125y \quad + \quad 0,2z \quad = \quad 270.$$

Similarly, we require (for blue and green colours respectively) that

$$\begin{array}{ccccccc} 0,25x & + & 0,75y & + & 0,2z & = & 270, \\ 0,25x & + & 0,125y & + & 0,6z & = & 270. \end{array}$$

From the first equation $x = 540 - 0,25y - 0,4z$. Substitute for x into the second and third equations to obtain two linear equations of two variables $2,75y + 0,4z = 540$ and $0,25y + 2z = 540$. From the second of these we express $z = 270 - 0,125y$ and substitute into the first one we obtain $2,7y = 432$, that is, $y = 160$. Therefore $z = 270 - 0,125 \cdot 160 = 250$ and hence $x = 540 - 0,25 \cdot 160 + 0,4 \cdot 250 = 400$.

An alternative approach is to deduce consequences from the given equations by a sequence of adding them or multiplying them by non-zero scalars. This is easily handled in the matrix notation (which we met when solving equations with two variables in the previous chapter already). The first row of the matrix consists of coefficients of the variables in the first equation, second of the coefficients in the second equation and third of the coefficients in the third. Therefore the *matrix of the system* is

$$\begin{pmatrix} 0,5 & 0,125 & 0,2 \\ 0,25 & 0,75 & 0,2 \\ 0,25 & 0,125 & 0,6 \end{pmatrix},$$

The *extended matrix of the system* is obtained from the matrix of the system by inserting the column of the right-hand sides

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 $u = (a_1, \ldots, a_n)$ is multiplied by a scalar c by multiplying every element of the n-tuple u by c. Addition is defined coordinate-wise.

<center>BASIC VECTOR OPERATIONS</center>

$$\begin{aligned} u + v &= (a_1, \ldots, a_n) + (b_1, \ldots, b_n) \\ &= (a_1 + b_1, \ldots, a_n + b_n) \\ c \cdot u = c \cdot (a_1, \ldots, a_n) &= (c \cdot a_1, \ldots, c \cdot a_n). \\ cu = c(a_1, \ldots, a_n) &= (ca_1, \ldots, ca_n). \end{aligned}$$

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.

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 vector 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) $0 = (0, \ldots, 0) \in \mathbb{K}^n$. 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:

<center>VECTOR PROPERTIES</center>

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)	$1 \cdot v = v$

of the individual equations in the system:

$$\left(\begin{array}{ccc|c} 0,5 & 0,125 & 0,2 & 270 \\ 0,25 & 0,75 & 0,2 & 270 \\ 0,25 & 0,125 & 0,6 & 270 \end{array} \right)$$

By doing elementary row transformations sequentially (they all correspond to adding rows and multiplication by scalars with the equations, see 2.1.7) we can eliminate the variables in the equations, one by one:

$$\left(\begin{array}{ccc|c} 0,5 & 0,125 & 0,2 & 270 \\ 0,25 & 0,75 & 0,2 & 270 \\ 0,25 & 0,125 & 0,6 & 270 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 1 & 3 & 0,8 & 1\,080 \\ 1 & 0,5 & 2,4 & 1\,080 \end{array} \right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 2,75 & 0,4 & 540 \\ 0 & 0,25 & 2 & 540 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 11 & 1,6 & 2\,160 \\ 0 & 1 & 8 & 2\,160 \end{array} \right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 1 & 8 & 2\,160 \\ 0 & 11 & 1,6 & 2\,160 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 1 & 8 & 2\,160 \\ 0 & 0 & -86,4 & -21\,600 \end{array} \right).$$

By back substitution, we compute successively

$$z = \frac{-21\,600}{-86,4} = 250,$$

$$y = 2\,160 - 8 \cdot 250 = 160,$$

$$x = 540 - 0,4 \cdot 250 - 0,25 \cdot 160 = 400.$$

Thus it is necessary to mix 400 litres of reddish, 160 litres of bluish and 250 litres of greenish colour. □

2.A.2. Solve the system of simultaneous linear equations

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 3x_3 & = & 2, \\ 2x_1 & - & 3x_2 & - & x_3 & = & -3, \\ -3x_1 & + & x_2 & + & 2x_3 & = & -3. \end{array}$$

Solution. We write the system of equations in the form of the extended matrix of the system

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 2 & -3 & -1 & -3 \\ -3 & 1 & 2 & -3 \end{array} \right).$$

Every row of the matrix corresponds to one equation. As in the previous example, equivalent transformation of the equations correspond to the elementary row operations on the matrix and we use them to transform it into the row echelon form

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 2 & -3 & -1 & -3 \\ -3 & 1 & 2 & -3 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -7 & -7 & -7 \\ 0 & 7 & 11 & 3 \end{array} \right) \sim$$

$$\sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 4 & -4 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{array} \right).$$

First we subtracted from the second row twice the first row, and to the third row we added three times the first row. Then

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 and 1.1.5, 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 $\mathbb{Z}^n$, $(\mathbb{Z}_k)^n$, $n = 1, 2, 3, \ldots$.

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

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

where $a_{ij} \in \mathbb{K}$ for all $1 \le i \le m$, $1 \le j \le n$. For a matrix A with elements a_{ij} we also use the notation $A = (a_{ij})$.

The vector $(a_{i1}, a_{i2}, \ldots, a_{in}) \in \mathbb{K}^n$ is called the (i-th) *row of the matrix* A, $i = 1, \ldots, m$. The vector $(a_{1j}, a_{2j}, \ldots, a_{mj}) \in \mathbb{K}^m$ is called the (j-th) *column of the matrix* A, $j = 1, \ldots, n$.

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

$$0 = \begin{pmatrix} 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix}$$

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 $(a_1, \ldots, a_n)$ and $(x_1, \ldots, x_n)$ their product

$$(a_1, \ldots, a_n) \cdot (x_1, \ldots, x_n) = a_1 x_1 + \cdots + a_n x_n.$$

This means, we multiply the corresponding coordinates of the vectors and sum the results.

we added the second row to the third row and multiplied the second row by $-1/4$. Now we restore the system of equations

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 3x_3 & = & 2, \\ & & x_2 & + & x_3 & = & 1, \\ & & & & x_3 & = & -1. \end{array}$$

We see immediately that $x_3 = -1$. If we substitute $x_3 = -1$ into the equation $x_2 + x_3 = 1$, we obtain $x_2 = 2$. Then by substituting $x_3 = -1$, $x_2 = 2$ into the first equation, we obtain $x_1 = 1$. $\square$

Systems of linear equations can be written in matrix notation. But is it an advantage, when we can solve the systems even without speaking about matrices? Yes it is, we can handle the equations more conceptually. We can easily decide how many solutions a system has. It is much more efficient in computer assisted computations. Thus we shall get familiar with various operations which can be done with matrices. As we have seen in previous examples, equivalent operations with linear equations correspond to elementary row (column) transformations. Further we have seen that transforming a matrix into a row echelon form, a process called Gaussian elimination, see 2.1.7), solves the system very easily. We demonstrate this on some examples, where we will see that a system can have infinitely many solutions or no solution at all.

2.A.3. Solve a system of linear equations

$$\begin{array}{rcrcrcr} 2x_1 & - & x_2 & + & 3x_3 & = & 0, \\ 3x_1 & + & 16x_2 & + & 7x_3 & = & 0, \\ 3x_1 & - & 5x_2 & + & 4x_3 & = & 0, \\ -7x_1 & + & 7x_2 & + & -10x_3 & = & 0. \end{array}$$

Solution. Because the right-hand side of all equations is zero (such a case is called a homogeneous system) we work with the matrix of the system only. We find the solution by transforming the matrix into the row echelon form using elementary row transformations. These correspond to changing the order of equations, multiplying an equation by a non-zero number and addition of multiples of equations. Furthermore, we can always go back and forth between the matrix notation and the original system notation with variables x_i. We obtain:

$$\begin{pmatrix} 2 & -1 & 3 \\ 3 & 16 & 7 \\ 3 & -5 & 4 \\ -7 & 7 & -10 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix}.$$

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 $(x_1, \ldots, x_n)$ (called the *vector of variables*, or *vector variable*) and the known vectors of coordinates $(a_{i1}, \ldots, a_{in})$.

The vector of variables can be also seen as a column in a matrix of the type $n/1$, and similarly the values $b_1, \ldots, b_n$ can be seen as a vector u, and that is again a single column of the matrix of the type $n/1$. Our system of equations can then be formally written as $A \cdot x = u$ as follows:

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

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:

<center>MATRIX PRODUCT</center>

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

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}, \text{ for arbitrary } 1 \le i \le m, 1 \le k \le q.$$

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

$$\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}.$$

From there we see that the second, third and fourth equations are multiples of the equation $7x_2 + x_3 = 0$. We continue:

$$\begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 7 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

Considered as equations, the last two are redundant, and we are left with just

$$\begin{array}{rcrcrcl} 2x_1 & - & x_2 & + & 3x_3 & = & 0, \\ & & 7x_2 & + & x_3 & = & 0 \end{array}$$

We substitute for the variable x_3 a parameter $t \in \mathbb{R}$ and express

$$x_2 = -\frac{1}{7}x_3 = -\frac{1}{7}t \quad \text{a} \quad x_1 = \frac{1}{2}(x_2 - 3x_3) = -\frac{11}{7}t.$$

If we now substitute $t = -7s$, we obtain the result in a simple form

$$(x_1,\, x_2,\, x_3) = (11s,\, s,\, -7s),\quad s \in \mathbb{R}.$$

The whole system has infinitely many solutions. $\qquad\square$

2.A.4. Find all solutions of the system of linear equations

$$\begin{array}{rcrcrcrcr} 3x_1 & & & + & 3x_3 & - & 5x_4 & = & -8, \\ x_1 & - & x_2 & + & x_3 & - & x_4 & = & -2, \\ -2x_1 & - & x_2 & + & 4x_3 & - & 2x_4 & = & 0, \\ 2x_1 & + & x_2 & - & x_3 & - & x_4 & = & -3. \end{array}$$

Solution. The corresponding extended matrix of the system is

$$\left(\begin{array}{cccc|c} 3 & 0 & 3 & -5 & -8 \\ 1 & -1 & 1 & -1 & -2 \\ -2 & -1 & 4 & -2 & 0 \\ 2 & 1 & -1 & -1 & -3 \end{array}\right).$$

By changing the order of rows (equations) we obtain

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & -8 \end{array}\right),$$

which we transform into the row echelon form:

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & -8 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & -3 & 6 & -4 & -4 \\ 0 & 3 & 0 & -2 & -2 \end{array}\right)$$

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 3 & -3 & -3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The system has thus infinitely many solutions, because we have three equations in four variables. These three equations have exactly one solution for any choice for the variable

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

$$E = (\delta_{ij}) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

is called the *unit matrix*, or alternatively, the *identity matrix*. The numbers δ_{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:

$$A \cdot B = \left(\sum_j a_{ij} \cdot b_{jk}\right), \quad B \cdot C = \left(\sum_k b_{jk} \cdot c_{kl}\right),$$

$$(A \cdot B) \cdot C = \left(\sum_k \left(\sum_j a_{ij}b_{jk}\right)c_{kl}\right) = \left(\sum_{j,k} a_{ij}b_{jk}c_{kl}\right),$$

$$A \cdot (B \cdot C) = \left(\sum_j a_{ij}\left(\sum_k b_{jk}c_{kl}\right)\right) = \left(\sum_{j,k} a_{ij}b_{jk}c_{kl}\right).$$

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:

$$A \cdot E = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & 1 \end{pmatrix} = A$$

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

$x_4 \in \mathbb{R}$. Thus for x_4 we substitute the parameter $t \in \mathbb{R}$ and go back from the matrix notation to the system of equations

$$
\begin{array}{rcrcrcrcr}
x_1 & - & x_2 & + & x_3 & - & t & = & -2, \\
 & & 3x_2 & - & 3x_3 & + & t & = & 1, \\
 & & & & 3x_3 & - & 3t & = & -3.
\end{array}
$$

From the last equation we have $x_3 = t - 1$. Substituting for x_3 into the second equation gives

$$3x_2 - 3t + 3 + t = 1, \quad \text{that is,} \quad x_2 = \frac{1}{3}\left(2t - 2\right).$$

Finally, using the first equation, we have

$$x_1 - \frac{1}{3}\left(2t - 2\right) + t - 1 - t = -2, \quad \text{tj.} \quad x_1 = \frac{1}{3}\left(2t - 5\right).$$

The set of solutions can be written (for $t = 3s$) in the form

$$\left\{(x_1, x_2, x_3, x_4) = \left(2s - \tfrac{5}{3}, 2s - \tfrac{2}{3}, 3s - 1, 3s\right), s \in \mathbb{R}\right\}.$$

We return to the extended matrix of the system and transform it further by using the row transformations in order to have (still in the row echelon form) the first non-zero number of every row (the so-called *pivot*) equal to one and that all the other numbers in the column of the pivot are zero. We have

$$
\left(
\begin{array}{cccc|c}
1 & -1 & 1 & -1 & -2 \\
0 & 3 & -3 & 1 & 1 \\
0 & 0 & 3 & -3 & -3 \\
0 & 0 & 0 & 0 & 0
\end{array}
\right)
$$

$$
\sim
\left(
\begin{array}{cccc|c}
1 & -1 & 1 & -1 & -2 \\
0 & 1 & -1 & 1/3 & 1/3 \\
0 & 0 & 1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0
\end{array}
\right)
$$

$$
\sim
\left(
\begin{array}{cccc|c}
1 & -1 & 0 & 0 & -1 \\
0 & 1 & 0 & -2/3 & -2/3 \\
0 & 0 & 1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0
\end{array}
\right)
$$

$$
\sim
\left(
\begin{array}{cccc|c}
1 & 0 & 0 & -2/3 & -5/3 \\
0 & 1 & 0 & -2/3 & -2/3 \\
0 & 0 & 1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0
\end{array}
\right),
$$

because first we have multiplied the second and the third row by $1/3$, then we have added the third row to the second and its (-1)-multiple to the first. Finally we have added the second row to the first. From the last matrix we easily obtain the result

$$
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}
=
\begin{pmatrix} -5/3 \\ -2/3 \\ -1 \\ 0 \end{pmatrix}
+ t
\begin{pmatrix} 2/3 \\ 2/3 \\ 1 \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}.
$$

Free variables are those whose columns do not contain any pivot (in our case there is no pivot in the fourth column, that is, the fourth variable is free and we use it as a parameter). $\square$

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

$$A \cdot (B + C) = \left(\sum_j a_{ij}(b_{jk} + c_{jk})\right)$$

$$= \left(\left(\sum_j a_{ij}b_{jk}\right) + \left(\sum_j a_{ij}c_{jk}\right)\right) = A \cdot B + A \cdot C$$

$$(B + C) \cdot D = \left(\sum_k (b_{jk} + c_{jk})d_{kl}\right)$$

$$= \left(\left(\sum_k b_{jk}d_{kl}\right) + \left(\sum_k c_{jk}d_{kl}\right)\right) = B \cdot D + C \cdot D.$$

As we have seen in 1.5.4, two matrices of dimension two do not necessarily commute: for example

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

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

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!) $\square$

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,

$$A \cdot (B \cdot C) = (A \cdot B) \cdot C$$
$$A \cdot (B + C) = A \cdot B + A \cdot C,$$

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*.

2.A.5. Determine the solutions of the system of equations

$$
\begin{array}{rcrcrcrcr}
3x_1 & & & + & 3x_3 & - & 5x_4 & = & 8, \\
x_1 & - & x_2 & + & x_3 & - & x_4 & = & -2, \\
-2x_1 & - & x_2 & + & 4x_3 & - & 2x_4 & = & 0, \\
2x_1 & + & x_2 & - & x_3 & - & x_4 & = & -3.
\end{array}
$$

Solution. Note that the system of equations in this exercise differs from the system of equations in the previous exercise only in the value 8 (instead of -8) on the right-hand side. If we do the same row transformations as in the previous exercise, we obtain

$$
\begin{pmatrix}
3 & 0 & 3 & -5 & | & 8 \\
1 & -1 & 1 & -1 & | & -2 \\
-2 & -1 & 4 & -2 & | & 0 \\
2 & 1 & -1 & -1 & | & -3
\end{pmatrix}
\sim
\begin{pmatrix}
1 & -1 & 1 & -1 & | & -2 \\
2 & 1 & -1 & -1 & | & -3 \\
-2 & -1 & 4 & -2 & | & 0 \\
3 & 0 & 3 & -5 & | & 8
\end{pmatrix}
$$

$$
\sim
\begin{pmatrix}
1 & -1 & 1 & -1 & | & -2 \\
0 & 3 & -3 & 1 & | & 1 \\
0 & 0 & 3 & -3 & | & -3 \\
0 & 0 & 3 & -3 & | & 13
\end{pmatrix}
\sim
\begin{pmatrix}
1 & -1 & 1 & -1 & | & -2 \\
0 & 3 & -3 & 1 & | & 1 \\
0 & 0 & 3 & -3 & | & -3 \\
0 & 0 & 0 & 0 & | & 16
\end{pmatrix},
$$

where the last operation was subtracting the third row from the fourth. From the fourth equation $0 = 16$ follows that the system has no solutions. Let us emphasize than whenever we obtain an equation of the form $0 = a$ for some $a \neq 0$ (that is, zero row on the left side and non-zero number after the vertical bar) when doing the row transformation, the system has no solutions. □

You can find more exercises for systems of systems of linear equations on the page

Now we are going to manipulate with matrices to get more familiar with their properties.

2.A.6. Matrix multiplication. Note that, in order to be able to multiply two matrices, the necessary and sufficient condition is that the first matrix has the same number of columns as the number of rows of the second matrix. The number of rows of the resulting matrix is then given by the number of rows of the first matrix, the number of columns then equals the number of columns of the second matrix.
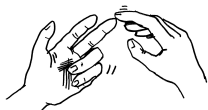
i) $\begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 5 & 4 \end{pmatrix},$

ii) $\begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 7 \end{pmatrix},$

iii) $\begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 2 & 1 \\ 1 & 1 & -2 & -3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$

$= \begin{pmatrix} 12 & 7 & 1 & -5 \\ 3 & 0 & 5 & 4 \end{pmatrix},$

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) $$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}.$$

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

$$
A \cdot x = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = u
$$

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,

iv) $\begin{pmatrix} 1 & 3 & 1 \\ -2 & 2 & -1 \\ 3 & 1 & -4 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3 \\ -3 \end{pmatrix} = \begin{pmatrix} 7 \\ 7 \\ 18 \end{pmatrix}$,

v) $\begin{pmatrix} 1 & 3 & -3 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 & 3 \\ 3 & 2 & 1 \\ 1 & -1 & -4 \end{pmatrix} = \begin{pmatrix} 7 & 7 & 18 \end{pmatrix}$,

vi) $\begin{pmatrix} 1 & 2 & -2 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = (-2)$.

Remark. Parts i) and ii) in the previous exercise show that multiplication of square matrices is not commutative in general. In part iii) we see that if we can multiply two rectangular matrices, then it is possible only in one of the orders. In parts iv) and v) note that $(A \cdot B)^T = B^T \cdot A^T$.

2.A.7. Let
$$A = \begin{pmatrix} 4 & 0 & -5 \\ 2 & 7 & 15 \\ 2 & 7 & 13 \end{pmatrix}, \quad B = \begin{pmatrix} 7 & 2 & 0 \\ 0 & 0 & 3 \\ 0 & -19 & \sqrt{13} \end{pmatrix}.$$

Can the matrix A be transformed into B using only elementary row transformations (we say then that such matrices are row equivalent)?

Solution. Both matrices are row equivalent with the three-dimensional identity matrix. It is easy to see that row equivalence on the set of all matrices of given type is indeed an equivalence relation. Thus the matrices A and B are row equivalent. □

2.A.8. Find a matrix B for which the matrix $C = B \cdot A$ is in row echelon form, where
$$A = \begin{pmatrix} 3 & -1 & 3 & 2 \\ 5 & -3 & 2 & 3 \\ 1 & -3 & -5 & 0 \\ 7 & -5 & 1 & 4 \end{pmatrix}.$$

Solution. If we multiply the matrix A successively from the left by elementary matrices (consider what elementary row transformations does it correspond to)

$$E_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$E_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -7 & 0 & 0 & 1 \end{pmatrix},$$

$$E_5 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_6 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

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 our ring of coordinates is an integral domain.

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 a 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 $k = 1, \ldots, j$, 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

$$\begin{pmatrix} 0 & \ldots & 0 & a_{1j} & \ldots & \ldots & \ldots & a_{1m} \\ 0 & \ldots & 0 & 0 & \ldots & a_{2k} & \ldots & a_{2m} \\ \vdots & & & & & & & \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & a_{lp} & \ldots \\ \vdots & & & & & & & \end{pmatrix}.$$

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:

$$E_7 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -4 & 0 & 1 \end{pmatrix}, \quad E_8 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

we obtain

$$B = E_8 E_7 E_6 E_5 E_4 E_3 E_2 E_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1/12 & -5/12 & 0 \\ 1 & -2/3 & 1/3 & 0 \\ 0 & -4/3 & -1/3 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & -3 & -5 & 0 \\ 0 & 1 & 9/4 & 1/4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

$\square$

2.A.9. Complex numbers as matrices.

Consider the set of matrices $C = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, a, b \in \mathbb{R} \right\}$. Note that C is closed under addition and matrix multiplication, and further show that the mapping $f : C \to \mathbb{C}$, $\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \mapsto a + bi$ satisfies $f(M + N) = f(M) + f(N)$ and $f(M \cdot N) = f(M) \cdot f(N)$ (on the left-hand sides of the equations we have addition and multiplication of matrices, on the right-hand sides we have addition and multiplication of complex numbers). Thus the set C along with multiplication and addition can be seen as the field $\mathbb{C}$ of complex numbers. The mapping f is called an isomorphism (of fields). Thus for instance we have

$$\begin{pmatrix} 3 & 5 \\ -5 & 3 \end{pmatrix} \cdot \begin{pmatrix} 8 & -9 \\ 9 & 8 \end{pmatrix} = \begin{pmatrix} 69 & 13 \\ -13 & 69 \end{pmatrix},$$

which corresponds to $(3 + 5i) \cdot (8 - 9i) = 69 - 13i$.

2.A.10. Solve the equations for matrices

$$\begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix} \cdot X_1 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad X_2 \cdot \begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Solution. Clearly the unknowns X_1 and X_2 must be matrices of the type 2×2 (in order for the products to be defined and that the result is a matrix of the type 2×2). Set

$$X_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$$

and multiply out the matrices in the first given equation. We obtain

$$\begin{pmatrix} a_1 + 3c_1 & b_1 + 3d_1 \\ 3a_1 + 8c_1 & 3b_1 + 8d_1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

(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 $i = 2, \ldots$, 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 nonzero 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. Matrix of elementary row transformations.

Let us now restrict ourselves to fields of scalars $\mathbb{K}$, that is, every nonzero 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)

$$\begin{pmatrix} \ddots & & & & \\ & 0 & \ldots & 1 & \\ & \vdots & \ddots & \vdots & \\ & 1 & \ldots & 0 & \\ & & & & \ddots \end{pmatrix} \begin{matrix} \\ \leftarrow i\text{-th row} \\ \\ \leftarrow j\text{-th row} \\ \\ \end{matrix}$$

that is,

$$
\begin{array}{rcrcrcrcl}
a_1 & & & + & 3c_1 & & & = & 1, \\
& & b_1 & & & + & 3d_1 & = & 2, \\
3a_1 & & & + & 8c_1 & & & = & 3, \\
& & 3b_1 & & & + & 8d_1 & = & 4.
\end{array}
$$

By adding a (-3)-multiple of the first equation with the third equation we obtain $c_1 = 0$ and then $a_1 = 1$. Analogously, by adding a (-3)-multiple of the second equation to the fourth equation we obtain $d_1 = 2$ and then $b_1 = -4$. Thus we have

$$
X_1 = \begin{pmatrix} 1 & -4 \\ 0 & 2 \end{pmatrix}.
$$

We can find the values a_2, b_2, c_2, d_2 by a different approach. If A is a square matrix, we write A^{-1} to denote its inverse, so that $A \cdot A^{-1} = A^{-1} \cdot A = E$, the unit matrix) It is easy to check that

$$
\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},
$$

which holds for any numbers $a, b, c, d \in \mathbb{R}$ provided $ad - bc \neq 0$. (This is easy to derive; it also directly follows from formula 1 in 2.2.11). We calculate

$$
\begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix}^{-1} = \begin{pmatrix} -8 & 3 \\ 3 & -1 \end{pmatrix}.
$$

Multiplying the given equations by this matrix from the right gives

$$
X_2 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} -8 & 3 \\ 3 & -1 \end{pmatrix},
$$

and thus

$$
X_2 = \begin{pmatrix} -2 & 1 \\ -12 & 5 \end{pmatrix}.
$$

□

2.A.11. Solve the matrix equation

$$
X \cdot \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 4 & -6 \\ 2 & 1 \end{pmatrix}.
$$

○

2.A.12. Computing the inverse matrix. 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 $\left(A^T \cdot B \right)^{-1}$.

Solution. We find the inverse by the following method: write next to each other the matrix A and the unit matrix. Then use elementary row transformations so that the sub-matrix A changes into the unit matrix. This will change the original unit sub-matrix to A^{-1}. We obtain

(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):

$$
i\text{-th row and }j\text{-th column} \rightarrow \begin{pmatrix}
\ddots & & & & \\
& 1 & & & \\
& & \ddots & & \\
& 1 & & 1 & \\
& & & & \ddots
\end{pmatrix}
$$

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*

$$
P \cdot A \cdot Q = \begin{pmatrix}
1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & & & & \\
0 & \cdots & 1 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & & & & &
\end{pmatrix}.
$$

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. □

$$\begin{pmatrix} 4 & 3 & 2 & | & 1 & 0 & 0 \\ 5 & 6 & 3 & | & 0 & 1 & 0 \\ 3 & 5 & 2 & | & 0 & 0 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & -2 & 0 & | & 1 & 0 & -1 \\ 5 & 6 & 3 & | & 0 & 1 & 0 \\ 3 & 5 & 2 & | & 0 & 0 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & -2 & 0 & | & 1 & 0 & -1 \\ 0 & 16 & 3 & | & -5 & 1 & 5 \\ 0 & 11 & 2 & | & -3 & 0 & 4 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & -2 & 0 & | & 1 & 0 & -1 \\ 0 & 5 & 1 & | & -2 & 1 & 1 \\ 0 & 11 & 2 & | & -3 & 0 & 4 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & -2 & 0 & | & 1 & 0 & -1 \\ 0 & 5 & 1 & | & -2 & 1 & 1 \\ 0 & 1 & 0 & | & 1 & -2 & 2 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & | & 3 & -4 & 3 \\ 0 & 0 & 1 & | & -7 & 11 & -9 \\ 0 & 1 & 0 & | & 1 & -2 & 2 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & | & 3 & -4 & 3 \\ 0 & 1 & 0 & | & 1 & -2 & 2 \\ 0 & 0 & 1 & | & -7 & 11 & -9 \end{pmatrix}.$$

In the first step we subtracted from the first row the third row, in the second step we added a (-5)-multiple of the first to the second row and added a (-3)-multiple of the first row to the third row, in the third step we subtracted from the second row the third row, in the fourth step we added a (-2)-multiple of the second row to the third row, in the fifth step we added a (-5)-multiple of the third row to the second row and added a 2-multiple of the third row to the first row, and in the last step we changed the second and the third row. We have obtained the result

$$A^{-1} = \begin{pmatrix} 3 & -4 & 3 \\ 1 & -2 & 2 \\ -7 & 11 & -9 \end{pmatrix}.$$

Note that when calculating the matrix A^{-1} we did not have to cope with fractions thanks to the suitably chosen row transformations. Although we could carry on similarly when doing the next exercise, that is, B^{-1}, we will rather do the more obvious row transformations. We have

$$\begin{pmatrix} 1 & 0 & 1 & | & 1 & 0 & 0 \\ 3 & 3 & 4 & | & 0 & 1 & 0 \\ 2 & 2 & 3 & | & 0 & 0 & 1 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & 0 & 1 & | & 1 & 0 & 0 \\ 0 & 3 & 1 & | & -3 & 1 & 0 \\ 0 & 2 & 1 & | & -2 & 0 & 1 \end{pmatrix} \sim$$

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 $PA = E$, 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. But it will be useful to understand their core meaning right

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 3 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1/3 & 0 & -2/3 & 1 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & \frac{1}{3} & -1 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & -\frac{2}{3} & 1 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 2 & -3 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & \frac{1}{3} & 0 & -\frac{2}{3} & 1 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 2 & -3 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 0 & -2 & 3 \end{pmatrix},$$

that is,

$$B^{-1} = \begin{pmatrix} 1 & 2 & -3 \\ -1 & 1 & -1 \\ 0 & -2 & 3 \end{pmatrix}.$$

Using the identity

$$\left(A^T \cdot B\right)^{-1} = B^{-1} \cdot \left(A^T\right)^{-1} = B^{-1} \cdot \left(A^{-1}\right)^T$$

and the knowledge of the inverse matrices computed before, we obtain

$$\left(A^T \cdot B\right)^{-1} = \begin{pmatrix} 1 & 2 & -3 \\ -1 & 1 & -1 \\ 0 & -2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 & -7 \\ -4 & -2 & 11 \\ 3 & 2 & -9 \end{pmatrix}$$

$$= \begin{pmatrix} -14 & -9 & 42 \\ -10 & -5 & 27 \\ 17 & 10 & -49 \end{pmatrix}.$$

2.A.13. Compute the inverse of the matrix

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 2 & -2 & 1 \\ 5 & -5 & 2 \end{pmatrix}.$$

2.A.14. Calculate A^5 and A^{-3}, if

$$A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.A.15. Compute the inverse of the matrix

$$\begin{pmatrix} 8 & 3 & 0 & 0 & 0 \\ 5 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 & 5 \end{pmatrix}.$$

2.A.16. Determine whether there exists an inverse of the matrix

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

If yes, then compute C^{-1}.

now. A linear combination of rows of a matrix $A = (a_{ij})$ of type m/n is understood as an expression of the form

$$c_1 u_{i_1} + \cdots + c_k u_{i_k},$$

where c_i are scalars, $u_j = (a_{j1}, \ldots, a_{jn})$ are rows of the matrix A. Similarly, we can consider linear combinations of columns by replacing the above rows u_j by the columns $u_j = (a_{1j}, \ldots, a_{mj})$.

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.

The previous results about the Gaussian elimination can be now interpreted as follows: the number of non-zero "steps" in the row (column) echelon form is always equal to the number of linearly independent rows (columns) of the matrix. Let E_h be the matrix from the theorem 2.1.9 with h ones on the diagonals and 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

$$E_h = P \cdot A \cdot Q, \ E_{h'} = P' \cdot A \cdot Q'.$$

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

$$P'' \cdot E_{h'} \cdot Q'' = E_h.$$

In the product $P'' \cdot E_{h'}$ there will be more zero rows in the bottom part of the echelon matrix than we see in E_h and we must be able to reach E_h using only elementary column transformations. This is clearly not possible, because the zero rows remain zero there.

Therefore 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. This number is called the *rank of the matrix* and we denote it by $h(A)$. We have the following theorem:

Theorem. *Let A be a matrix of type m/n over a field of scalars $\mathbb{K}$. The matrix A has the same 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.A.17. Compute A^{-1}, if

(a) $A = \begin{pmatrix} 1 & i \\ -i & 3 \end{pmatrix}$, while i is the imaginary unit

(b) $A = \begin{pmatrix} 1 & -5 & -3 \\ -1 & 5 & 4 \\ -1 & 6 & 2 \end{pmatrix}$.

2.A.18. Find the inverse to the $n \times n$ matrix $(n > 1)$

$$A = \begin{pmatrix} 2-n & 1 & \cdots & 1 & 1 \\ 1 & 2-n & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & \ddots & \ddots & 2-n & 1 \\ 1 & 1 & \cdots & 1 & 2-n \end{pmatrix}.$$

Solution. You can try for small n ($n = 2, 3, 4$), which is easy to compute with the known algorithm, and then guess the general form.

$$A^{-1} = \frac{1}{n-1} \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}.$$

□

We have already encountered systems of linear equations at the beginning of the chapter. Now we will deal with them in more detail. We use the inverse matrix to assist in computing the solution to the system of linear equations. Note that we do the same computation as before. To express the variables is the same as to bring the matrix of the system with equivalent transformation to the identity matrix and that is the same as to multiply the matrix of the system with the inverse matrix.

2.A.19. Participants of a trip. There were 45 participants of a two-day bus trip. On the first day, the fee for a watchtower visit was €30 for an adult, €16 for a child and €24 for a senior. The total fee for the first day was €1 116. On the second day, the fee for a bus with a palace and botanical garden tour was €40 for an adult, €24 for a child and €34 for a senior. The total fee for the second day was €1 542. How many adults, children and seniors were there among the participants?

Solution. Introduce the variables

x for the „number of adults“;

y for the „number of children“;

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

$$A : \mathbb{K}^n \to \mathbb{K}^m, \quad x \mapsto A \cdot x.$$

By the distributivity of matrix multiplication, it is clear how the linear combinations of vectors are mapped using such mappings:

$$A \cdot (a\,x + b\,y) = a\,(A \cdot x) + b\,(A \cdot y).$$

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 $A : \mathbb{K}^n \to \mathbb{K}^n$.

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.

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 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.

z for the „number of seniors";

There were 45 participants, therefore

$$x \;+\; y \;+\; z \;=\; 45.$$

The fees for the first and second days respectively imply that

$$\begin{aligned} 30x &+& 16y &+& 24z &=& 1\,116, \\ 40x &+& 24y &+& 34z &=& 1\,542. \end{aligned}$$

We write the system of three linear equations in the matrix notation as

$$\begin{pmatrix} 1 & 1 & 1 \\ 30 & 16 & 24 \\ 40 & 24 & 34 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 45 \\ 1\,116 \\ 1\,542 \end{pmatrix}.$$

We compute

$$\begin{pmatrix} 1 & 1 & 1 \\ 30 & 16 & 24 \\ 40 & 24 & 34 \end{pmatrix}^{-1} = \frac{1}{6} \begin{pmatrix} 16 & 5 & -4 \\ 30 & 3 & -3 \\ -40 & -8 & 7 \end{pmatrix}.$$

Hence the solution is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 16 & 5 & -4 \\ 30 & 3 & -3 \\ -40 & -8 & 7 \end{pmatrix} \cdot \begin{pmatrix} 45 \\ 1\,116 \\ 1\,542 \end{pmatrix}$$

$$= \frac{1}{6} \begin{pmatrix} 132 \\ 72 \\ 66 \end{pmatrix} = \begin{pmatrix} 22 \\ 12 \\ 11 \end{pmatrix},$$

expressed in words, there were 22 adults, 12 children and 11 seniors. $\square$

The latter approach is particularly efficient if we have to solve several systems with the same matrix on the left hand side but different values on the right hand side.

But what if the matrix of the system is not invertible? Then we cannot use the inverse matrix for solving the system. Such a system cannot have a single solution. As the reader may have noticed above, a system of linear equations either has no solution, has one solution or has infinitely many solutions, depending on one or more free parameters (for instance, it cannot have exactly two solutions). We should have also noticed when dealing with equations with two variables in the previous section, that the space of the solutions is either a vector space (in the case when the right-hand side of the system is zero, we speak of a *homogeneous system* of linear equations) or an affine space, see 4.1.1 (in the case when the right-hand side of at least one of the equations is non-zero, we speak of a *non-homogeneous system* of linear equations).

We can recognize all the possibilities from the rank of the matrices, i.e. the number of nonzero rows left in the row-echelon form.

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 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 volume 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 volume. 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 $X = \{1, 2, \ldots, n\}$, the permutation can be written by putting the resulting ordering into a table:

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}.$$

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 σ 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 σ 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):

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

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*:

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 of linear equations

$$\begin{array}{rrrrrrrrr}
x_1 & + & x_2 & + & x_3 & - & 2x_4 & = & 4, \\
-3x_1 & - & 2x_2 & - & x_3 & - & x_4 & = & 5, \\
 & + & 2x_2 & & & + & x_4 & = & 1, \\
x_1 & - & 4x_2 & + & x_3 & - & 2x_4 & = & 3
\end{array}$$

Determine also all solutions of the system

$$\begin{array}{rrrrrrrrr}
x_1 & + & x_2 & + & x_3 & - & 2x_4 & = & 0, \\
-3x_1 & - & 2x_2 & - & x_3 & - & x_4 & = & 0, \\
 & + & 2x_2 & & & + & x_4 & = & 0, \\
x_1 & - & 4x_2 & + & x_3 & - & 2x_4 & = & 0
\end{array}$$

and of the system

$$\begin{array}{rrrrrrr}
x_1 & - & 3x_2 & & & = & 1, \\
x_1 & - & 2x_2 & + & 2x_3 & = & -4, \\
x_1 & - & x_2 & & & = & 1, \\
-2x_1 & - & x_2 & + & x_3 & = & -2.
\end{array}$$

Solution. Transforming the matrix to the row-echelon form, we check that the rank is four. (The rank cannot exceed the number of rows or columns). The first of the three given system is given by the extended matrix

$$\left(\begin{array}{rrrr|r} 1 & 1 & 1 & -2 & 4 \\ -3 & -2 & -1 & -1 & 5 \\ 0 & 2 & 0 & 1 & 1 \\ 1 & -4 & 1 & -2 & 3 \end{array} \right).$$

But the left-hand side is exactly A^T and thus we can get the column-echelon form the same way as before. In particular, the columns of the matrix are linearly indepent and the rank is maximal, i.e. four again. Therefore there exists a matrix $\left(A^T\right)^{-1}$ and the system has a unique solution

$$(x_1,\, x_2,\, x_3,\, x_4)^T = \left(A^T\right)^{-1} \cdot (4, 5, 1, 3)^T.$$

The second of the systems 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 given by

$$(x_1,\, x_2,\, x_3,\, x_4) = (0, 0, 0, 0).$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{array}{l} a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} \\ -a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} \end{array}$$

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

$$|A| = \sum_{\sigma \in \Sigma_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where Σ_n is the set of all possible permutations over $\{1, \ldots, n\}$ and the symbol sgn for a permutation σ, called the parity of σ, will be described below. Each of the expressions

$$\operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

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 $a, b \in X = \{1, \ldots, n\}$ forms an *inversion in the permutation σ, if $a < b$ and $\sigma(a) > \sigma(b)$*. A permutation σ is called *even* or *odd*, if it contains an even or odd number of inversions, respectively.

Thus, the *parity* $\operatorname{sgn} \sigma$ *of the permutation σ is* $(-1)^{\text{number of inversions}}$ and we denote it by $\operatorname{sgn}(\sigma)$. 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 $X = \{1, 2, \ldots, n\}$ 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 σ there is such a sequence starting with σ.

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 $\sigma(1) = a_1, \ldots, \sigma(n) = a_n$. 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

The third system is given by the extended matrix

$$\left(\begin{array}{ccc|c} 1 & -3 & 0 & 1 \\ 1 & -2 & 2 & -4 \\ 1 & -1 & 0 & 1 \\ -2 & -1 & 1 & -2 \end{array}\right),$$

which is the matrix A (only the last column is given after the vertical bar). If we try to simplify the matrix into the row echelon form, we must obtain a row

$$\left(\begin{array}{ccc|c} 0 & 0 & 0 & a \end{array}\right), \quad \text{where} \quad a \neq 0.$$

We know, that 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 4). This system thus has no solution. □

For further examples see 2.E.7

B. Determinants
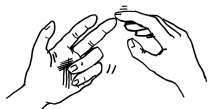
In order to be able to define the key object of the matrix calculus, the determinant, we must deal with permutations (bijections of a finite set) and their parities.

We shall use the two-row notation for permutations (see 2.2.1). In the first row we list all elements of the given set, and every column then corresponds to a pair (preimage, image) in the given permutation. Because a permutation is a bijection, the second row is indeed a permutation (ordering) of the first row, in accordance with the definition from combinatorics.

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 transpositions.

Solution. We first decompose the permutation into a product of independent cycles. Start with the first element 1 and look on the second row to see what the image of 1 is. It is 3. Now look on the column that starts with 3, and see that the image of 3 is 6, and so on. Continue until we again reach the starting element 1. We obtain the following sequence of elements, which map to each other under the given permutation:

$$1 \mapsto 3 \mapsto 6 \mapsto 9 \mapsto 2 \mapsto 1.$$

The mapping which maps elements in such a manner is called a cycle (see 2.2.3) which we denote by $(1, 3, 6, 9, 2)$.

Now choose any element not contained in the obtained cycle. With the same procedure as with 1, we obtain the cycle $(4, 7, 5, 8)$. From the method is clear that the result does not depend on the first obtained cycle. Each element from the set

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

$$(a_1, \ldots, a_i, a_{i+1}, \ldots, a_n),$$

containing r inversions. Then in the ordering

$$(a_1, \ldots, a_{i+1}, a_i, \ldots, a_n)$$

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. □

We found that applying a transposition changes the parity of a permutation and any ordering of numbers $\{1, 2, \ldots, n\}$ can be obtained through transposing of neighbouring elements. Therefore we have proven

Corollary. *On every finite set $X = \{1, \ldots, n\}$ 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 $\sigma, \eta : X \to X$ we have

$$\mathrm{sgn}(\sigma \circ \eta) = \mathrm{sgn}(\sigma) \cdot \mathrm{sgn}(\eta)$$

and also

$$\mathrm{sgn}(\sigma^{-1}) = \mathrm{sgn}(\sigma).$$

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.

CYCLES

A permutation σ over the set $X = \{1, \ldots, n\}$ is called a *cycle* of length k, if we can find elements $a_1, \ldots, a_k \in X$, $2 \leq k \leq n$ such that $\sigma(a_i) = a_{i+1}$, $i = 1, \ldots, k - 1$, while $\sigma(a_k) = a_1$, and other elements in X are fixed-points of σ. Cycles of length two are transpositions.

Every permutation is a composition of cycles. Cycles of even length have parity -1, cycles of odd length have parity 1.

PROOF. The last claim has yet to be proved. Fix a permutation σ 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 σ (notice σ^{-1} means the inverse bijection to σ). Clearly, it is an equivalence relation

($\{1, 2, \ldots, 9\}$) appears in one of the obtained cycles, we can thus write:

$$\sigma = (1, 3, 6, 9, 2) \circ (4, 7, 5, 8),$$

or

$$\sigma = (4, 7, 5, 8) \circ (1, 3, 6, 9, 2),$$

since independent cycles commute. For cycles the decomposition into transpositions is simple, we have

$$(1, 3, 6, 9, 2) = (1, 3) \circ (3, 6) \circ (6, 9) \circ (9, 2) =$$
$$(1, 3)(3, 6)(6, 9)(9, 2).$$

Thus we obtain:

$$\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8).$$

$\square$

Remark. The minimal number of transpositions in the decomposition of a permutation is obtained by carrying out exactly the procedure as above. That is, first decompose the permutation into the independent cycles, then the cycles canonically into the transpositions. Thus the found decomposition is the decomposition into the minimal number of transpositions.

Note also that the operation $\circ$ is a composition of mappings, thus it is necessary to carry out the composition "backwards", as we are used to in composition of mappings. Applying the given composition of transposition for instance on the element two we can successively write:

$$[(1, 3)(3, 6)(6, 9)(9, 2)](2) =$$
$$[(1, 3)(3, 6)(6, 9)]((9, 2)(2)) =$$
$$[(1, 3)(3, 6)(6, 9)](9) = [(1, 3)(3, 6)](6) = (1, 3)(3) = 1,$$

thus the mapping indeed maps the element 2 on the element 1 (it is actually just the cycle $(1, 3, 6, 9, 2)$ written in a different way). When writing a composition of permutations, we often omit the sign "$\circ$" and speak of the product of permutations.

When writing the cycle we write only the elements on which the cycle (that is, the mapping) nontrivially acts (that is, the element is mapped to some other element). Fixed-points of the cycle are not listed. Thus it is necessary to know on which set do we consider the given cycle (mostly it will be clear from the context). The cycle $c = (4, 7, 5, 8)$ from the previous example is thus a mapping (permutation), which, in the two-row notation, looks like this

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 2 & 3 & 7 & 8 & 6 & 5 & 4 & 9 \end{pmatrix}.$$

If the original permutation has some fixed-points they do not appear in the cycle decomposition.

(check it carefully!). Because X is a finite set, for some ℓ it must be that $\sigma^\ell(x) = x$. If we pick one equivalence class $\{x, \sigma(x), \ldots, \sigma^{\ell-1}(x)\} \subset X$ 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. $\square$

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 σ 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 $\operatorname{sgn}\sigma a_{ij}a_{1\sigma(1)} \ldots^\wedge \ldots a_{n\sigma(n)}$ where the hat denotes the omition 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 σ. Clearly, $i - 1$ changes on the domain and $j - 1$ changes on the target do the job (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.

84

Note further that the notation $(1, 2, 3)$ gives the same cycle as for instance $(2, 3, 1)$ or $(3, 1, 2)$. But the notation $(1, 3, 2)$ is a different cycle.

2.B.2. Determine the parity of the following permutations:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 1 & 6 & 7 & 8 & 9 & 5 & 4 & 2 \end{pmatrix},$$
$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 1 & 5 & 3 \end{pmatrix}.$$

Solution. According to our definition (see 2.2.2) we compute the number of inversions of σ: we go sequentially through the second row in the two-row notation and for every number k there we count the number of numbers which are smaller than k and are located after k in the second row. It is not hard to see that the number of inversions in a given permutation is exactly the number of pairs "larger before smaller" in the second row. For σ we compute (stepping through the second row): after three there is one and two, thus we add 2; after one there is no smaller number and we add 0; after six there is five, four and two, thus we add 3, similarly for seven, eight and nine, for five we add 2, for four we add 1 and for two nothing. Thus we have 17 inversions in total and thus the permutation is odd.

But we can compute the parity of σ otherwise. The theorem 2.2.2 implies that the parity of a permutation is given by the parity of the number of transpositions in its decomposition (this number is, unlike the number of transposition in an arbitrary decomposition, always the same)

The previous exercise gives us $\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8)$. There are seven transpositions in the decomposition, thus the permutation is indeed odd.

Alternatively we can decompose τ into either a product of three transpositions (using the cycle decomposition):

$$\tau = (1, 2, 4)(3, 6) = (1, 2)(2, 4)(3, 6),$$

or we count the number of inversions in τ: $1+2+3+0+1 = 7$. Either way we find that τ is an odd permutation.

In general, as soon as the decomposition to cycles is ready, we may just count the lengths of the cycles, since each cycle including k elements is clearly built of $k - 1$ transpositions and thus contributes $(-1)^{k-1}$ to the parity. $\square$

For the following exercises, recall how to compute determinants of the type 2×2 ($a_{11} \cdot a_{22} - a_{12} \cdot a_{21}$) and 3×3 (Sarrus rule), see 2.2.1.

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 omiting 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 containg 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*.

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.

PROOF. (1) The terms of determinants $|A|$ and $|A^T|$ are in bijective correspondence, where the term $\mathrm{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$ corresponds the following A^T term (notice it does not depend on the order of scalars)

$$\mathrm{sgn}(\sigma) a_{\sigma(1)1} \cdot a_{\sigma(2)2} \cdots a_{\sigma(n)n} =$$
$$= \mathrm{sgn}(\sigma) a_{1\sigma^{-1}(1)} \cdot a_{2\sigma^{-1}(2)} \cdots a_{n\sigma^{-1}(n)},$$

and we have to ensure that this member has the correct sign. But the parities of σ and σ^{-1} are the same, and so this is really a term in the determinant $|A^T|$ and the first claim is proved.

2.B.3. Compute the determinant of the following matrices

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

Solution. The determinant of the first matrix is $1 \cdot 1 - 2 \cdot 2 = -3$.

As for the second matrix, according to the Sarrus rule we just have to enumerate the expression
$1\cdot(-1)\cdot2+2\cdot2\cdot3+3\cdot1\cdot2-3\cdot(-1)\cdot3-1\cdot2\cdot2-1\cdot2\cdot2 = 17.$

We can also bring the matrix into the row echelon form and then multiply the numbers on the diagonal but we have to remember that a multiplication of a row with a scalar changes the determinant of the matrix by the same multiple. Interchanging two rows changes the sign of the determinant of the matrix.

$$\begin{vmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 2 & 2 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -1 \\ 0 & -4 & -7 \end{vmatrix} = \frac{1}{-4}\cdot\frac{1}{3}\cdot\begin{vmatrix} 1 & 2 & 3 \\ 0 & 12 & 4 \\ 0 & -12 & -21 \end{vmatrix}$$
$$= -\frac{1}{12}\cdot\begin{vmatrix} 1 & 2 & 3 \\ 0 & 12 & 4 \\ 0 & 0 & -17 \end{vmatrix}$$

We finish with an upper triangular matrix. The determinant of such matrices is the product of the numbers on the main diagonal. So the result is $-\frac{1}{12}(1 \cdot 12 \cdot (-17)) = 17$.

We can see, that using the Sarrus rule is quicker.

For the third matrix we have
$1\cdot0\cdot1+1\cdot0\cdot1+1\cdot0\cdot(-2)-1\cdot0\cdot(-2)-1\cdot1\cdot1-1\cdot0\cdot0 = -1.$

$\square$

It is important to realize, that Sarrus rule can be used for matrices 3×3 only. For higher dimension matrices you can either bring the matrix to the row echelon form (where you have to take in to account rules 2.2.5) or use the Laplace expansion (see 2.2.9).

2.B.4. Compute the determinant of the matrix

$$\begin{pmatrix} 1 & 3 & 5 & 6 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{pmatrix}.$$

Solution. We compute this in two ways. First, convert the matrix to row echelon form. We can use already known elementary transformations,

(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. 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).

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

If A is in the row echelon form then
$$|A| = a_{11}a_{22}\cdots a_{nn}.$$
The previous theorem gives an effective method for computing determinants using the Gauss elimination method, 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 $|A| = a_{11}a_{22}\cdots a_{kk}|B|$.

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

$$
\begin{vmatrix} 1 & 3 & 5 & 6 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{vmatrix} = - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 5 & 6 \\ 0 & 1 & 2 & 1 \end{vmatrix} = - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 2 & 4 & 4 \\ 0 & 1 & 2 & 1 \end{vmatrix}
$$

$$
= - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 4 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{vmatrix} = 2.
$$

Note, that we have interchanged the rows twice in the course of computation.

The other way of computing the determinant is by cofactor expansion along the first column (the one with the greatest number (one) of zeroes). Successively we obtain

$$
\begin{vmatrix} 1 & 3 & 5 & 6 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{vmatrix} = 1 \cdot \begin{vmatrix} 2 & 2 & 2 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} - 1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} +
$$

$$
1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 2 & 2 & 2 \\ 1 & 2 & 1 \end{vmatrix} \overset{\text{using the Sarrus rule}}{=} -2 - 2 + 6 = 2.
$$

$\square$

2.B.5. Compute the determinant of the matrix

$$
\begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 3 & 0 & 3 \\ 4 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}
$$

Solution. We notice, that the last (fifth) row contains four zeros (as well as the second column). It is the most, we can find in a row or a column in the matrix, thus it will be advantageous to use Laplace theorem (2.3.10) and compute the determinant via expasion along the fifth row or second column. We present the expansion via fifth row:

$$
\begin{vmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 3 & 0 & 3 \\ 4 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{vmatrix} = 0 \cdot \begin{vmatrix} 0 & 1 & 0 & 1 \\ 2 & 0 & 2 & 0 \\ 0 & 3 & 0 & 3 \\ 0 & 0 & 4 & 4 \end{vmatrix} - 0 \cdot \begin{vmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 3 & 0 & 3 \\ 4 & 0 & 4 & 4 \end{vmatrix}
$$

$$
+ 0 \cdot \begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 4 & 0 & 4 & 4 \end{vmatrix} - 0 \cdot \begin{vmatrix} 1 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 3 \\ 4 & 0 & 0 & 4 \end{vmatrix} + 5 \cdot \begin{vmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 \\ 4 & 0 & 0 & 4 \end{vmatrix}
$$

$$
= 5 \cdot \begin{vmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 \\ 4 & 0 & 0 & 4 \end{vmatrix} = 5 \cdot 2 \cdot \begin{vmatrix} 1 & 1 & 0 \\ 0 & 3 & 0 \\ 4 & 0 & 4 \end{vmatrix} = 120,
$$

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.

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 shall work with field of scalars now.

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 $b = (b_1, \ldots, b_n)$. 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 $x = (x_1, \ldots, x_n)$ 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. 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 $|A|^{-1}$ exists. 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_ℓ-multiples of all the other ℓ-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 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 $x_1, \ldots, x_n$ and their images $A \cdot x_1, \ldots, A \cdot x_n$.

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:

where we have used the expansion along the second column in the second step and computed the determinant of the 3×3 matrix directly using the Sarrus rule.

Another option is to try to expand the determinant along several rows, exploiting vanishing of many sub-determinants there. For example, we may use the last two rows. Clearly there might be only two non-zero sub-determinants built from this row there. Thus the entire determinant must be (notice that choosing two lines and two columns always leads to the plus sign in the definition of the algebraic complement, see 2.3.10)

$$\begin{vmatrix} 4 & 4 \\ 0 & 5 \end{vmatrix} \cdot \begin{vmatrix} 0 & 1 & 0 \\ 2 & 0 & 2 \\ 0 & 3 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 4 \\ 0 & 5 \end{vmatrix} \cdot \begin{vmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{vmatrix} =$$
$$= 20 \cdot 0 + 20 \cdot 6 = 120$$

$\square$

2.B.6. Find all the values of a such that

$$\begin{vmatrix} a & 1 & 1 & 1 \\ 0 & a & 1 & 1 \\ 0 & 1 & a & 1 \\ 0 & 0 & 0 & -a \end{vmatrix} = 1.$$

For complex a give either its algebraic or polar form.

Solution. We compute the determinant by expanding the first row of the matrix:

$$D = \begin{vmatrix} a & 1 & 1 & 1 \\ 0 & a & 1 & 1 \\ 0 & 1 & a & 1 \\ 0 & 0 & 0 & -a \end{vmatrix} = a \cdot \begin{vmatrix} a & 1 & 1 \\ 1 & a & 1 \\ 0 & 0 & -a \end{vmatrix}.$$

Expand further using the last row:

$$D = a \cdot (-a) \begin{vmatrix} a & 1 \\ 1 & a \end{vmatrix} = -a^2(a^2 - 1).$$

We conclude that $a^4 - a^2 + 1 = 0$. Substituting $t = a^2$ we have $t^2 - t + 1$ with roots $t_1 = \frac{1+i\sqrt{3}}{2} = \cos(\pi/3) + i\sin(\pi/3)$, $t_2 = \frac{1-i\sqrt{3}}{2} = \cos(\pi/3) - i\sin(\pi/3) = \cos(-\pi/3) + i\sin(-\pi/3)$, from where we obtain four possible values for the parameter a: $a_1 = \cos(\pi/6) + i\sin(\pi/6) = \sqrt{3}/2 + i/2$, $a_2 = \cos(7\pi/6) + i\sin(7\pi/6) = -\sqrt{3}/2 - i/2$, $a_3 = \cos(-\pi/6) + i\sin(-\pi/6) = \sqrt{3}/2 - i/2$, $a_4 = \cos(5\pi/6) + i\sin(5\pi/6) = -\sqrt{3}/2 + i/2$.

Alternatively, we can multiply by $a^2 + 1$ to obtain

$$a^6 + 1 = (a^2 + 1)(a^4 - a^2 + 1) = 0.$$

The equation $a^6 = -1$ has six (complex) solutions given by $a = \cos\varphi + i\sin\varphi$ where $\varphi = \pi/6 + k\pi/3 = (2k+1)\pi/6$, $k = 0, 1, 2, 3, 4, 5$. Of these, we must discard the two choices $k = 1$, and $k = 4$, since these choices solve $a^2 + 1 = 0$ and

---

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.

Let $A = (a_{ij})$ be a matrix of the type m/n and let $1 \leq i_1 < \ldots < i_k \leq m$, $1 \leq j_1 < \ldots < j_l \leq n$ be fixed natural numbers. Then the matrix

$$M = \begin{pmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \ldots & a_{i_1 j_\ell} \\ \vdots & & & \vdots \\ a_{i_k j_1} & a_{i_k j_2} & \ldots & a_{i_k j_\ell} \end{pmatrix}$$

of the type k/ℓ is called a *submatrix of the matrix A* determined by the rows $i_1, \ldots, i_k$ and columns $j_1, \ldots, j_\ell$. 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. 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

$$(-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_l} \cdot |M^*|$$

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*.

not $a^4 - a^2 + 1 = 0$. We conclude that $a = \cos\varphi + i\sin\varphi$ where $\varphi = (2k+1)\pi/6$, $k = 0, 2, 3$, or 5. $\qquad\square$

2.B.7. Vandermonde determinant. Prove the formula for the Vandermonde determinant, that is, the determinant of the Vandermonde matrix:

$$V_n = \begin{vmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2 & \dots & x_n^{n-1} \end{vmatrix} = \prod_{1 \le i < j \le n}(x_j - x_i),$$

where $x_1, \dots, x_n \in \mathbb{R}$ and on the right-hand side of the equation there is the product of all terms $x_j - x_i$ where $j > i$.

Solution. We proceed by induction on n. From technical reasons we work with the transposed Vandermonde matrix (it has the same determinant). By subtracting the first row from all other rows and then expanding the first column we obtain

$V_n(x_1, x_2, \dots, x_n)$

$$= \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix}$$

$$= \begin{vmatrix} x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix}.$$

If we take out $x_{i+1} - x_1$ from the i-th row for $i \in \{1, 2, \dots, n-1\}$, we obtain

$$V_n(x_1, x_2, \dots, x_n) = (x_2 - x_1)\cdots(x_n - x_1)$$

$$\cdot \begin{vmatrix} 1 & x_2 + x_1 & \dots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \dots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix}.$$

By subtracting from every column (starting with the last and ending with the second) x_1-multiple of the previous column, we obtain

$$\begin{vmatrix} 1 & x_2 + x_1 & \dots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \dots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix}$$

$$= \begin{vmatrix} 1 & x_2 & \dots & x_2^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-2} \end{vmatrix}.$$

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 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 $i_1 < i_2 < \cdots < i_k$ and columns $j_1 < \cdots < j_k$. Then using $(i_1 - 1) + \cdots + (i_k - k)$ exchanges of neighbouring rows and $(j_1 - 1) + \cdots + (j_k - k)$ exchanges of neighbouring columns in A we can transform this submatrix M into a 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 $\alpha = \sum_{h=1}^k (i_h + j_h) - 2(1 + \cdots + k)$. 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 $k!(n-k)!$ 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 $(-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_l} \cdot |M| \cdot |M^*|$ 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 $(-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_l} \cdot |M| \cdot |M^*|$ of minors of the order k chosen among the fixed rows with their algebraic complements.*

Therefore

$$V_n(x_1, x_2, \ldots, x_n)$$
$$= (x_2 - x_1) \cdots (x_n - x_1) \; V_{n-1}(x_2, \ldots, x_n).$$

Because it is clear that

$$V_2(x_{n-1}, x_n) = x_n - x_{n-1},$$

it follows by induction that

$$V_n(x_1, x_2, \ldots, x_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i).$$

Note that the determinant is non-zero whenever the numbers $x_1, \ldots, x_n$ are mutually distinct. □

Remark. Another (more beautiful?) proof of the formula can be found in 5.1.5.

2.B.8. Find whether or not the matrix

$$\begin{pmatrix} 3 & 2 & -1 & 2 \\ 4 & 1 & 2 & -4 \\ -2 & 2 & 4 & 1 \\ 2 & 3 & -4 & 8 \end{pmatrix}$$

is invertible.

Solution. The matrix is invertible (that is, there is an inverse matrix) whenever we can transform it by elementary row transformations into the unit matrix. That is equivalent for instance to the property that it has non-zero determinant. That we can compute using the Laplace Theorem (2.3.10) by expanding for instance the first row:

$$\begin{vmatrix} 3 & 2 & -1 & 2 \\ 4 & 1 & 2 & -4 \\ -2 & 2 & 4 & 1 \\ 2 & 3 & -4 & 8 \end{vmatrix} = 3 \cdot \begin{vmatrix} 1 & 2 & -4 \\ 2 & 4 & 1 \\ 3 & -4 & 8 \end{vmatrix}$$

$$-2 \cdot \begin{vmatrix} 4 & 2 & -4 \\ -2 & 4 & 1 \\ 2 & -4 & 8 \end{vmatrix} + (-1) \cdot \begin{vmatrix} 4 & 1 & -4 \\ -2 & 2 & 1 \\ 2 & 3 & 8 \end{vmatrix}$$

$$- 2 \cdot \begin{vmatrix} 4 & 1 & 2 \\ -2 & 2 & 4 \\ 2 & 3 & -4 \end{vmatrix}$$

$$= 3 \cdot 90 - 2 \cdot 180 + (-1) \cdot 110 - 2 \cdot (-100) = 0,$$

that is, the given matrix is not invertible. □

2.B.9. Solve the system from 2.A.2 using the Cramer rule (see 2.2.6).

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)matrices A, B, and so on).

$$H = \begin{pmatrix} A & 0 \\ -E & B \end{pmatrix} = \begin{pmatrix} a_{11} & \ldots & a_{1n} & 0 & \ldots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} & 0 & \ldots & 0 \\ -1 & & 0 & b_{11} & \ldots & b_{1n} \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & b_{n1} & \ldots & b_{nn} \end{pmatrix}$$

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 with zeros in the bottom right corner. We obtain

$$K = \begin{pmatrix} a_{11} & \ldots & a_{1n} & c_{11} & \ldots & c_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} & c_{n1} & \ldots & c_{nn} \\ -1 & & 0 & 0 & \ldots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & 0 & \ldots & 0 \end{pmatrix}.$$

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.

Solution. We just plug in the values to the rule:

$$x_1 = \frac{\begin{vmatrix} 2 & 2 & 3 \\ -3 & -3 & -1 \\ -3 & 1 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = 1, \ x_2 = \frac{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & -1 \\ -3 & -3 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = 2$$

$$x_3 = \frac{\begin{vmatrix} 1 & 2 & 2 \\ 2 & -3 & -3 \\ -3 & 1 & -3 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = -1.$$

$\square$

2.B.10. Find the algebraically adjoint matrix and the inverse of the matrix

$$A = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 3 & 0 & 4 \\ 5 & 0 & 6 & 0 \\ 0 & 7 & 0 & 8 \end{pmatrix}.$$

Solution. The adjoint matrix is

$$A^* = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}^T,$$

where 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 have

$$A_{11} = \begin{vmatrix} 3 & 0 & 4 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} = -24, \quad A_{12} = -\begin{vmatrix} 0 & 0 & 4 \\ 5 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = 0,$$

$$A_{13} = \begin{vmatrix} 0 & 3 & 4 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 20, \quad A_{14} = -\begin{vmatrix} 0 & 3 & 0 \\ 5 & 0 & 6 \\ 0 & 7 & 0 \end{vmatrix} = 0,$$

$$A_{21} = -\begin{vmatrix} 0 & 2 & 0 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} = 0, \quad A_{22} = \begin{vmatrix} 1 & 2 & 0 \\ 5 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = -32,$$

$$A_{23} = -\begin{vmatrix} 1 & 0 & 0 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 0, \quad A_{24} = \begin{vmatrix} 1 & 0 & 2 \\ 5 & 0 & 6 \\ 0 & 7 & 0 \end{vmatrix} = -28,$$

$$A_{31} = \begin{vmatrix} 0 & 2 & 0 \\ 3 & 0 & 4 \\ 7 & 0 & 8 \end{vmatrix} = 8, \quad A_{32} = -\begin{vmatrix} 1 & 2 & 0 \\ 0 & 0 & 4 \\ 0 & 0 & 8 \end{vmatrix} = -0,$$

$$A_{33} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 7 & 8 \end{vmatrix} = -4, \quad A_{34} = -\begin{vmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 0 & 7 & 0 \end{vmatrix} = -0,$$

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 a 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*

$$(1) \qquad AA^* = A^*A = |A| \cdot E.$$

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 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.

91

$$A_{41} = - \begin{vmatrix} 0 & 2 & 0 \\ 3 & 0 & 4 \\ 0 & 6 & 0 \end{vmatrix} = 0, \quad A_{42} = \begin{vmatrix} 1 & 2 & 0 \\ 0 & 0 & 4 \\ 5 & 6 & 0 \end{vmatrix} = -16,$$

$$A_{43} = - \begin{vmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 0 \end{vmatrix} = 0, \quad A_{44} = \begin{vmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 5 & 0 & 6 \end{vmatrix} = -12.$$

By substitution we obtain

$$A^* = \begin{pmatrix} -24 & 0 & 20 & 0 \\ 0 & -32 & 0 & 28 \\ 8 & 0 & -4 & 0 \\ 0 & 16 & 0 & -12 \end{pmatrix}^T$$

$$= \begin{pmatrix} -24 & 0 & 8 & 0 \\ 0 & -32 & 0 & 16 \\ 20 & 0 & -4 & 0 \\ 0 & 28 & 0 & -12 \end{pmatrix}.$$

We compute the inverse matrix A^{-1} from the relation $A^{-1} = |A|^{-1} \cdot A^*$. The determinant of the matrix A is (expanding the first row) equal to

$$|A| = \begin{vmatrix} 1 & 0 & 2 & 0 \\ 0 & 3 & 0 & 4 \\ 5 & 0 & 6 & 0 \\ 0 & 7 & 0 & 8 \end{vmatrix} = \begin{vmatrix} 3 & 0 & 4 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} + 2 \begin{vmatrix} 0 & 3 & 4 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 16.$$

By substitution, we obtain

$$A^{-1} = \begin{pmatrix} -3/2 & 0 & 1/2 & 0 \\ 0 & -2 & 0 & 1 \\ 5/4 & 0 & -1/4 & 0 \\ 0 & 7/4 & 0 & -3/4 \end{pmatrix}.$$

$\square$

C. Vector spaces and linear mappings

Typical properties of vector spaces (met already in the plane or three dimensional space) can be observed in many other situations. We illustrate this by examples.

2.C.1. Vector space – yes or no? Decide whether following sets form a vector space over the field of real numbers:

i) The set of solutions of the system

$$x_1 + x_2 + \cdots + x_{98} + x_{99} + x_{100} = 100x_1,$$

$$x_1 + x_2 + \cdots + x_{98} + x_{99} = 99x_1,$$

$$x_1 + x_2 + \cdots + x_{98} = 98x_1,$$

$$\vdots$$

$$x_1 + x_2 = 2x_1.$$

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 the homogeneous system $A \cdot x = 0$, that is

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

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.

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.

ii) The set of solutions of the equation

$$x_1 + x_2 + \cdots + x_{100} = 0$$

iii) The set of solutions of the equation

$$x_1 + 2x_2 + 3x_3 + \cdots + 100x_{100} = 1.$$

iv) The set of all real (or complex) sequences. (Real or complex sequence is a mapping $f : \mathbb{N} \to \mathbb{R}$ or $f : \mathbb{N} \to \mathbb{C}$. The image of number n is then called n-th member of the sequence, we usually denote it by lower index, say a_n.)

v) The set of solutions of a homogeneous difference equation.

vi) The set of solutions of a non-homogeneous difference equation.

vii) $\{f : \mathbb{R} \to \mathbb{R} | f(1) = f(2) = c, c \in \mathbb{R}\}$

Solution. We check the properties of a vector space, see 2.3.1. Actually all we have to do is to check whether the given sets are closed to linear combinations of it's elements. Then all the axioms of a vector space are satisfied.

i) Yes. They all are real multiples of the vector $\underbrace{(1, 1, 1 \ldots, 1)}_{100 \text{ ones}}$. A sum of two multiples of the same vector is again a multiple of the vector. The reverse vector is again a multiple of the vcetor and all other axioms are trivially satisfied. By the way, the solution space is thus a vector space of dimension 1, see also 2.3.7.

ii) Yes. It is a space of dimension 99 (corresponds to the number of free parameters of the solution). In general the set of all solutions of any system of homogeneous linear equations forms a vector space.

iii) No. For instance, taking twice the solution $x_1 = 1, x_i = 0, i = 2, \ldots 100$ we do not obtain a solution. But the set of solutions forms an affine space (see 4.1.1).

iv) Yes. The set of all real or complex sequences clearly forms a real (complex) vector space. Adding the sequences and scalar multiplication is defined term-wise, where it is clearly the vector space of all real (complex) numbers.

v) Yes. In order to show that the set of sequences which satisfy given difference homogeneous equation it is enough to show that it is closed under addition and real number multiplication (as the set of all real sequences is a vector space, as we know). Consider two sequences $(x_j)_{j=0}^{\infty}$

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 69 hold.

Recall our simple notational convention: scalars are usually denoted by letters from the beginning of the alphabet, that is, $a, b, c, \ldots$, 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, ℓ, 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) $\left(\sum_{i=1}^{n} a_i \right) \cdot \left(\sum_{j=1}^{m} u_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \cdot u_j.$

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 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 $u = 1 \cdot u = (p^{-1} \cdot p) \cdot u = p^{-1} \cdot 0 = 0$. □

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:

and $(y_j)_{j=0}^{\infty}$ satisfying the given 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

$$
a_n(x_{n+k} + y_{n+k}) + a_{n-1}(x_{n+k-1} + y_{n+k-1}) \\
+ \cdots + a_0(x_k + y_k) = 0,
$$

therefore 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 also $(ux_j)_{j=0}^{\infty}$, where $u \in \mathbb{R}$.

vi) No. The sum of two solutions of a non-homogeneous equation

$$
a_n x_{n+k} + a_{n-1} x_{n+k-1} + \cdots + a_0 x_k = c
$$
$$
a_n y_{n+k} + a_{n-1} y_{n+k-1} + \cdots + a_0 y_k = c, \; c \in \mathbb{R} - \{0\}
$$

satisfies the equation

$$
a_n(x_{n+k} + y_{n+k}) + a_{n-1}(x_{n+k-1} + y_{n+k-1}) \\
+ \cdots + a_0(x_k + y_k) = 2c,
$$

that is, it does not satisfy the original non-homogeneous equation. But the set of solutions forms an affine space, see 4.1.1.

vii) It is a vector space if and only if $c = 0$. If we take two functions f and g from the given set, then $(f + g)(1) = (f + g)(2) = f(1) + g(1) = 2c$. Thus if $f + g$ is to be a member of the given set, it must be that $(f + g)(1) = c$, therefore $2c = c$, hence $c = 0$.

$\square$

2.C.2. Find out, whether the set

$$
U_1 = \{(x_1, x_2, x_3) \in \mathbb{R}^3; \, |x_1| = |x_2| = |x_3|\}
$$

is a subspace of a vector space $\mathbb{R}^3$ and the set

$$
U_2 = \{ax^2 + c; \, a, c \in \mathbb{R}\}
$$

a subspace of the space of polynomials of degree at most 2.

Solution. The only property we have to check is whether the given subset is closed under linear combination of vectors in it, that is if it forms a vector space. The set U_1 is not a vector (sub)space. We can see that, for instance,

$$
(1, 1, 1) + (-1, 1, 1) = (0, 2, 2) \notin U_1.
$$

An expression of the form $a_1 v_1 + \cdots + a_k v_k$ is called a *linear combination* of vectors $v_1, \ldots, v_k \in V$.

A finite sequence of vectors $v_1, \ldots, v_k$ is called *linearly independent*, if the only zero linear combination is the one with all coefficients zero. That is, for any scalars $a_1, \ldots, a_k \in \mathbb{K}$, $a_1 v_1 + \cdots + a_k v_k = 0$ implies $a_1 = a_2 = \cdots = a_k = 0$. 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 $v_1, \ldots, v_k \in M$ 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.

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 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 $f : \mathbb{R} \to \mathbb{R}$ and define the addition and scalar multiplication like this: $(f + g)(x) = f(x) + g(x)$, $(a \cdot f)(x) = a \cdot f(x)$.

The set U_2 is a subspace (there is a clear identification with $\mathbb{R}^2$), because

$$\left(a_1 x^2 + c_1\right) + \left(a_2 x^2 + c_2\right) = \left(a_1 + a_2\right) x^2 + \left(c_1 + c_2\right),$$

$$k \cdot \left(ax^2 + c\right) = (ka) x^2 + kc$$

for all numbers $a_1, c_1, a_2, c_2, a, c, k \in \mathbb{R}$. $\qquad\square$

2.C.3. Determine whether or not the vectors $(1, 2, 3, 1)$, $(1, 0, -1, 1)$, $(2, 1, -1, 3)$ and $(0, 0, 3, 2)$ are linearly independent.

Solution. Because

$$\begin{vmatrix} 1 & 2 & 3 & 1 \\ 1 & 0 & -1 & 1 \\ 2 & 1 & -1 & 3 \\ 0 & 0 & 3 & 2 \end{vmatrix} = 10 \neq 0,$$

the given vectors are linearly independent. $\qquad\square$

2.C.4. Given arbitrary linearly independent vectors u, v, w, z in a vector space V, decide whether or not in V the vectors

$$u - 2v, \quad 3u + w - z, \quad u - 4v + w + 2z, \quad 4v + 8w + 4z$$

are linearly independent.

Solution. Considered vectors are linearly independent if and only if the vectors $(1, -2, 0, 0)$, $(3, 0, 1, -1)$, $(1, -4, 1, 2)$, $(0, 4, 8, 4)$ are linearly independent in $\mathbb{R}^4$. We have

$$\begin{vmatrix} 1 & -2 & 0 & 0 \\ 3 & 0 & 1 & -1 \\ 1 & -4 & 1 & 2 \\ 0 & 4 & 8 & 4 \end{vmatrix} = -36 \neq 0,$$

thus the vectors are linearly independent. $\qquad\square$

2.C.5. The vectors

$$(1, 2, 1), \quad (-1, 1, 0), \quad (0, 1, 1)$$

are linearly independent, and therefore together form a basis of $\mathbb{R}^3$ (for basis it is important to give an order of the vectors). Every three-dimensional vector is therefore some linear combination of them. What linear combination corresponds to the vector $(1, 1, 1)$, or equivalently, what are the coordinates of the vector $(1, 1, 1)$ in the basis formed by the given vectors?

Solution. We seek a, b, $c \in \mathbb{R}$ such that $a(1, 2, 1) + b(-1, 1, 0) + c(0, 1, 1) = (1, 1, 1)$. The equation must hold in every coordinate, so we have a system of three linear equations in three variables:

$$\begin{aligned} a - b &= 1 \\ 2a + b + c &= 1 \\ a + c &= 1, \end{aligned}$$

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 $\mathbb{R} \to \mathbb{R}$ or of all mappings $M \to V$ 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, $a, b \in \mathbb{K}$, $u, v \in \cap_{i \in I} W_i$. Then $a \cdot u + b \cdot v \in W_i$ for all $i \in I$. Hence $a \cdot u + b \cdot v \in \cap_{i \in I} W_i$.

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 span M.

We say that a set M *generates* the subspace span M, or that the elements of M are *generators* of the subspace span M.

We formulate a few simple claims about subspace generation:

Proposition. *For every nonempty set $M \subset V$, we have*

(1) span $M = \{a_1 \cdot u_1 + \cdots + a_k \cdot u_k; \ k \in \mathbb{N}, a_i \in \mathbb{K}, u_j \in M, j = 1, \ldots, k\}$;

(2) $M =$ span M *if and only if M is a vector subspace;*

(3) *if $N \subset M$ then* span $N \subset$ span M *is a vector subspace; the subspace* span $\emptyset$ *generated by the empty subspace is the trivial subspace $\{0\} \subset V$.*

PROOF. (1) The set of all linear combinations

$$a_1 u_1 + \cdots + a_k u_k$$

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 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. $\quad\square$

BASIS AND DIMENSION

A subset $M \subset V$ is called a *basis of the vector space V* if 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 *infinitely dimensional*. We write $\dim V = k$, $k \in \mathbb{N}$ or $k = \infty$.

whose solution gives us $a = \frac{1}{2}, b = -\frac{1}{2}, c = \frac{1}{2}$, thus we have

$$(1,1,1) = \frac{1}{2} \cdot (1,2,1) - \frac{1}{2} \cdot (-1,1,0) + \frac{1}{2} \cdot (0,1,1),$$

that is, the coordinates of the vector $(1,1,1)$ in the basis $((1,2,1),(-1,1,0),(0,1,1))$ are $(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$. $\square$

2.C.6. Determine all constants $a \in \mathbb{R}$ such that the polynomials $ax^2 + x + 2$, $-2x^2 + ax + 3$ and $x^2 + 2x + a$ are linearly dependent (in the vector space $P_3[x]$ of polynomials of one variable of degree at most three over real numbers).

Solution. In the basis 1, x, x^2 the coefficients of the given vectors (polynomials) are $(a,1,2)$, $(-2,a,3)$, $(1,2,a)$. Polynomials are linearly independent if and only if the matrix whose columns are given by the coordinates of the vectors has a rank lower than the number of the vectors. In this case the rank must be two or less. In the case of a square matrix, a rank less than the number of rows means that the determinant is zero. The condition for a thus reads

$$\begin{vmatrix} a & -2 & 1 \\ 1 & a & 2 \\ 2 & 3 & a \end{vmatrix} = 0,$$

that is, a is a root of the polynomial $a^3 - 6a - 5 = (a+1)(a^2 - a - 5)$, thus there are 3 such constants $a_1 = -1$, $a_2 = \frac{1+\sqrt{21}}{2}$, $a_3 = \frac{1-\sqrt{21}}{2}$. $\square$

2.C.7. 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 $x^2 + x + 1$.

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 $2 + i$ 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:

$$-\frac{1}{2}a - \frac{1}{2}b = 2$$
$$\frac{\sqrt{3}}{2}a - \frac{\sqrt{3}}{2}b = 1.$$

The solution gives us $a = -2 + \frac{\sqrt{3}}{3}$, $b = -2 - \frac{\sqrt{3}}{3}$, therefore the coordinates are $(-2 + \frac{1}{\sqrt{3}}, -2 - \frac{1}{\sqrt{3}})$. $\square$

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. 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

$$e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{K}^n, \quad i = 1, \ldots, n$$

(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 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.

2.C.8. Remark. As a perceptive reader may have spotted, the problem statement is not unambiguous – we are not given the order of the roots of the polynomial, thus we do not have the order of the basis vectors. The result is thus given up to the permutation of the coordinates.

We add a remark about rationalising the denominator, that is, removing the square roots from the denominator. The authors do not have a distinctive attitude whether this should always be done or not (Does $\frac{\sqrt{3}}{3}$ look better than $\frac{1}{\sqrt{3}}$?). In some cases the rationalising is undesirable: from the fraction $\frac{6}{\sqrt{35}}$ we can immediately spot that its value is a little greater than 1 (because $\sqrt{35}$ is just a little smaller than 6), while for the rationalised fraction $\frac{6\sqrt{35}}{35}$ we cannot spot anything. But in general the convention is to normalize.

2.C.9. Consider 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.C.10. For what values of the parameters $a, b, c \in \mathbb{R}$ are the vectors $(1, 1, a, 1)$, $(1, b, 1, 1)$, $(c, 1, 1, 1)$ linearly dependent? ◯

2.C.11. Let a vector space V be given along with a basis formed by the vectors u, v, w, z. Determine whether or not the vectors

$$u - 3v + z, \quad v - 5w - z, \quad 3w - 7z, \quad u - w + z$$

are linearly independent. ◯

2.C.12. Complete the vectors $1 - x^2 + x^3, 1 + x^2 + x^3, 1 - x - x^3$ to a basis of the space of polynomials of degree at most 3. ◯

2.C.13. Do the matrices

$$\begin{pmatrix} 1 & 0 \\ 1 & -2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 4 \\ 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} -5 & 0 \\ 3 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & -2 \\ 0 & 3 \end{pmatrix}$$

form a basis of the vector space of square two-dimensional matrices?

Solution. The four given matrices are as vectors in the space of 2×2 matrices linearly independent. It follows from the fact that the matrix

$$\begin{pmatrix} 1 & 1 & -5 & 1 \\ 0 & 4 & 0 & -2 \\ 1 & 0 & 3 & 0 \\ -2 & -1 & 0 & 3 \end{pmatrix}$$

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

$$A \cdot (x - y) = b - b = 0$$
$$A \cdot (x + z) = b + 0 = b.$$

Thus we can summarise in the form of the so called *Kronecker-Capelli theorem*[1]:

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 $V_1, \ldots, V_k \subset V$ 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 aggregate summands that belong to the same subspace and for a finite sum of k subspaces we obtain

$$V_1 + V_2 + \cdots + V_k = \{v_1 + \cdots + v_k; \ v_i \in V_i, i = 1, \ldots, k\}.$$

---

[1] 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.

is invertible (which is by the way equivalent to any of the following claims: its rank equals its dimension; it can be transformed into the unit matrix by elementary row transformations; it has the inverse matrix; it has a non-zero determinant (equal to 116); it stands for a system of homogeneous linear equations with only zero solution; every non-homogeneous linear system with left-hand side given by this matrix has a unique solution; the range of a linear mapping given by this matrix is a vector space of dimension 4 – this mapping is injective). □

2.C.14. In the vector space $\mathbb{R}^4$ we are given three-dimensional subspaces

$$U = \mathrm{span}\{u_1, u_2, u_3\}, \qquad V = \mathrm{span}\{v_1, v_2, v_3\},$$

while

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \ u_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \ u_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad v_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

$v_2 = (1, -1, 1, -1)^T$, $v_3 = (1, -1, -1, 1)^T$. Determine the dimension and find a basis of the subspace $U \cap V$.

Solution. The subspace $U \cap V$ contains exactly the vectors that can be obtained as a linear combinations of vectors u_i and also as a linear combination of vectors v_i. Thus we search for numbers $x_1, x_2, x_3, y_1, y_2, y_3 \in \mathbb{R}$ such that the following holds:

$$x_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} = y_1 \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + y_2 \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} + y_3 \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix},$$

that is, we are looking for a solution of a system

$$\begin{array}{rcrcrcrcrcr}
x_1 & + & x_2 & + & x_3 & = & y_1 & + & y_2 & + & y_3, \\
x_1 & + & x_2 & & & = & y_1 & - & y_2 & - & y_3, \\
x_1 & & & + & x_3 & = & -y_1 & + & y_2 & - & y_3, \\
& & x_2 & + & x_3 & = & -y_1 & - & y_2 & + & y_3.
\end{array}$$

Using matrix notation of this homogeneous system (and preserving the order of the variables) we have

$$\begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 0 & -1 & 1 & 1 \\ 1 & 0 & 1 & 1 & -1 & 1 \\ 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 0 & -1 & 0 & 2 & 2 \\ 0 & -1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

The sum $W = V_1 + \cdots + V_k \subset V$ is called the *direct sum* of subspaces if the intersection of any two is trivial, that is, $V_i \cap V_j = \{0\}$ for all $i \neq j$. We show that in such a case, every vector $w \in W$ can be written in a unique way as the sum

$$w = v_1 + \cdots + v_k,$$

where $v_i \in V_i$. Indeed, if we could simultaneously write w as $w = v_1' + \cdots + v_k'$, then

$$0 = w - w = (v_1 - v_1') + \cdots + (v_k - v_k').$$

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

$$W = V_1 \oplus \cdots \oplus V_k = \oplus_{i=1}^k V_i.$$

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 indepence of the number of basis elements on any choices.

A basis of a k-dimensional space will usually be denoted as a k-tuple $\underline{v} = (v_1 \ldots, v_k)$ 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 $(v_1, \ldots, v_n)$ is a basis of V, then the whole space V is the direct sum of the one-dimensional subspaces

$$V = \mathrm{span}\{v_1\} \oplus \cdots \oplus \mathrm{span}\{v_n\}.$$

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

$$w = x_1 v_1 + \cdots + x_n v_n.$$

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.

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 $V = \mathrm{span}\{v\}$ and $v \neq 0$,

$$\sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 1 & 1 & 1 & 1 & -1 \\ 0 & 0 & -1 & 0 & 2 & 2 \\ 0 & 0 & 1 & 3 & 1 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 1 & 1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

We obtain a solution

$x_1 = -2t,\ x_2 = -2s,\ x_3 = 2s + 2t,\ y_1 = -s - t,\ y_2 = s,$

$$y_3 = t, \quad t, s \in \mathbb{R}.$$

We obtain a general vector of the intersection by substituting

$$\begin{pmatrix} x_1 + x_2 + x_3 \\ x_1 + x_2 \\ x_1 + x_3 \\ x_2 + x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -2t - 2s \\ 2s \\ 2t \end{pmatrix}.$$

We see that

$$\dim U \cap V = 2, \quad U \cap V = \operatorname{span}\left\{ \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

$\square$

2.C.15. Let there be in $\mathbb{R}^3$ two vector spaces U and V generated by the vectors

$(1, 1, -3), (1, 2, 2) \quad \text{and} \quad (1, 1, -1), (1, 2, 1), (1, 3, 3),$

respectively. Determine the intersection of these two subspaces.

Solution. 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 $(1, 1, -1), (1, 2, 1), (1, 3, 3)$. It helps to consider first the geometry. Firstly, U is spanned by two linearly independent vectors. So U is a plane in $\mathbb{R}^3$. Next, V is spanned by three vectors. But these are linearly dependent since

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ -1 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \end{vmatrix} = 0.$$

So V is also a plane.

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 $V = \operatorname{span}\{v_1, \ldots, v_{n+1}\}$. If $v_1, \ldots, v_{n+1}$ are linearly independent, then they form a basis. If they are linearly dependent, there exists i such that

$$v_i = a_1 v_1 + \cdots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \cdots + a_{n+1} v_{n+1}.$$

Then $V = \operatorname{span}\{v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{n+1}\}$ and we can choose a basis, using the inductive assumption.

In remains to show that bases always have the same number of elements. Consider a basis $\underline{v} = (v_1, \ldots, v_n)$ of the space V and for an arbitrary nonzero vector u, consider

$$u = a_1 v_1 + \cdots + a_n v_n \in V$$

with $a_i \neq 0$ for some i. Then

$$v_i = \frac{1}{a_i} \big(u - (a_1 v_1 + \cdots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \cdots + a_n v_n) \big)$$

and therefore also $\operatorname{span}\{u, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n\} = V$.

We show that this is again a basis. For if adding u to the linearly independent vectors $v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n$ leads to a set of linearly dependent vectors, then

$$V = \operatorname{span}\{v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n\},$$

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 $(u, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ is again a basis of V.

Similarly, instead of one vector u, we can consider a linearly independent set $u_1, \ldots, u_k$. We will sequentially add $u_1, u_2, \ldots$, 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 $u_1, \ldots, u_\ell$ 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 $u_1, \ldots, u_\ell$ were nonzero, then it would mean that the vectors $u_1, \ldots, u_{\ell+1}$ 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 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*:

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 λ, μ. Similarly (x_1, x_2, x_3) lies in V, so $(x_1, x_2, x_3) = \alpha(1, 1, -1) + \beta(1, 2, 1) + \gamma(1, 3, 3)$ for scalars α, β, γ. When written in full, this is a set of six equations in eight unknowns. Solving these is possible but can be quite cumbersome. Some simplification is obtained as follows:

The first three equations, which describe U are

$$x_1 = \lambda + \mu$$
$$x_2 = \lambda + 2\mu$$
$$x_3 = -3\lambda + 2\mu$$

If we solve these three equations for the two "unknowns" λ and μ, (which in any case we do not want), or alternatively if we eliminate λ and μ, from these equations, we obtain the single equation $8x_1 - 5x_2 + x_3 = 0$ to replace the first three.

The second set of three equations, which describe V are

$$x_1 = \alpha + \beta + \gamma$$
$$x_2 = \alpha + 2\beta + \gamma$$
$$x_3 = -\alpha + \beta + 3\gamma$$

If we solve these three equations for the three "unknowns" α β and γ, (which in any case we do not want), or alternatively if we eliminate α β and γ, from these equations, we obtain the single equation $3x_1 - 2x_2 + x_3 = 0$ to describe V. Introducing the parameter t, it is straightforward write the solution as the line $(x_1, x_2, x_3) = t(3, 5, 1)$.

$\square$

Now we move to unions of vector spaces. There is a simple algorithm, how to chose the maximal linearly independent set of vectors out of a given set of vectors. Write the given vectors as columns in a matrix. Then tranform the matrix with row tranformation into the row echelon forms. The vectors, who correspond to the columns, where the "stairs" begin, form a maximal linearly independent set. To justify this, just think of the system of linear equations describing that a linear combination of the given vectors is zero. The matrix of the system is exactly the one described. If you put in the system only vectors corresponding to columns where are stairs, you get a system which can be transformed into a one in row echelon form with non-zero numbers on the diagonal and thus only solution of the systems are zeros, that is the vectors are linearly independet. Similarly, the system together with any

For every finite basis $\underline{v}$ of a vector space V and every set of linearly independent vectors u_i, $i = 1, \ldots, k$ in V we can find a subset of the basis vectors v_i 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 $(w_1 \ldots, w_t)$ 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 $(w_1, \ldots, w_t, u_{t+1} \ldots, u_r)$ for W_1 and to a basis $(w_1 \ldots, w_t, v_{t+1}, \ldots, v_s)$ for W_2. Vectors

$$w_1, \ldots, w_t, u_{t+1}, \ldots, u_r, v_{t+1} \ldots, v_s$$

clearly generate $W_1 + W_2$. We show that they are linearly independent. Let

$$a_1 w_1 + \cdots + a_t w_t + b_{t+1} u_{t+1} + \ldots$$
$$\cdots + b_r u_r + c_{t+1} v_{t+1} + \cdots + c_s v_s = 0.$$

Then necessarily

$$-(c_{t+1} \cdot v_{t+1} + \cdots + c_s \cdot v_s) =$$
$$= a_1 \cdot w_1 + \cdots + a_t \cdot w_t + b_{t+1} \cdot u_{t+1} + \cdots + b_r \cdot u_r$$

must belong to $W_2 \cap W_1$. This implies that

$$b_{t+1} = \cdots = b_r = 0,$$

since this is the way we have defined our bases. Then also

$$a_1 \cdot w_1 + \cdots + a_t \cdot w_t + c_{t+1} \cdot v_{t+1} + \cdots + c_s \cdot v_s = 0$$

of the vectors which correspond to "no stair" columns has a lower rank than the number of variables (coefficients of a linear combination), thus according to 2.3.5 has a nontrivial (non-zero) solution.

2.C.16. Determine the vector subspace (of the space $\mathbb{R}^4$) generated by the vectors $u_1 = (-1, 3, -2, 1)$, $u_2 = (2, -1, -1, 2)$, $u_3 = (-4, 7, -3, 0)$, $u_4 = (1, 5, -5, 4)$, by choosing a maximal set of linearly independent vectors u_i (that is, by choosing a basis).

Solution. Write the vectors u_i into the columns of a matrix and transform it using elementary row transformations. This way we obtain

$$\begin{pmatrix} -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \\ 1 & 2 & 0 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 4 & -4 & 5 \\ 0 & -7 & 7 & -7 \\ 0 & 3 & -3 & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} ① & 0 & 2 & 0 \\ 0 & ① & -1 & 0 \\ 0 & 0 & 0 & ① \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

And (according to the algorithm) it follows that the vectors corresponding to the columns with circled elements, namely vectors u_1, u_2 and u_4 form a maximal linearly independent set. □

Remark. Note, that the maximal set of linearly independent vectors is not unique. Unique is only the number of vectors in it (the dimension of the vector space generated by the given vectors). For example from vectors $(1, 0)$, $(0, 1)$, $(1, 1)$ you can pick any two to form a maximal linearly independent set, from vectors $(1, 0)$, $(2, 0)$, $(0, 1)$. This fact is also reflected in the algorithm, becouse it is independent of an order, in which you put the given vectors as columns in a matrix.

2.C.17. Find a basis of the subspace

$$U = \mathrm{span}\left\{ \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 1 & 2 \\ 3 & 4 \end{pmatrix}, \begin{pmatrix} -2 & -1 \\ 0 & 1 \\ 2 & 3 \end{pmatrix} \right\}$$

of the vector space of real matrices 3×2. Extend this basis to a basis of the whole space.

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. □

2.3.10. Examples. (1) $\mathbb{K}^n$ has (as a vector space over $\mathbb{K}$) dimension n. The n-tuple of vectors

$$((1, 0, \ldots, 0), (0, 1, \ldots, 0) \ldots, (0, \ldots, 0, 1))$$

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 $1, x, x^2, \ldots, x^m$.

The vector space of all polynomials $\mathbb{K}[x]$ has dimension ∞, but we can still find a basis (although infinite in size): $1, x, x^2, \ldots$.

(4) The vector space $\mathbb{R}$ over $\mathbb{Q}$ has dimension ∞. It does not have a countable basis.

(5) The vector space of all mappings $f : \mathbb{R} \to \mathbb{R}$ has also dimension ∞. It does not have any countable basis.

2.3.11. Vector coordinates. If we fix a basis $(v_1, \ldots, v_n)$ of a finite dimensional space V, then every vector $w \in V$ can be expressed as a linear combination $v = a_1 v_1 + \cdots + a_n v_n$ in a unique way. Indeed, assume that we can do it in two ways:

$$w = a_1 v_1 + \cdots + a_n v_n = b_1 v_1 + \cdots + b_n v_n.$$

Then

$$0 = (a_1 - b_1) \cdot v_1 + \cdots + (a_n - b_n) \cdot v_n$$

and thus $a_i = b_i$ for all $i = 1, \ldots, n$, because the vectors v_i are linearly independent. We have reached the concept of coordinates:

COORDINATES OF VECTORS

Definition. The coefficients of the unique linear combination expressing the given vector $w \in V$ in the chosen basis $\underline{v} = (v_1, \ldots, v_n)$ are called the *coordinates of the vector w* in this basis.

Whenever we speak about coordinates $(a_1, \ldots, a_n)$ of a vector w, which we express as a sequence, we must have a fixed ordering of the basis vectors $\underline{v} = (v_1, \ldots, v_n)$. Although we have defined the basis as a minimal set of generators, in reality we work with them as with sequences (that is, with ordered sets).

Solution. Recall that a basis of a subspace is a set of linearly independent vectors which generate given subspace. By writing the entries of the matrices in a row, we can consider the matrices as vectors in R^6. In this way, the four given matrices can be identified with the rows of the matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ -1 & 0 & 1 & 2 & 3 & 4 \\ -2 & -1 & 0 & 1 & 2 & 3 \end{pmatrix}.$$

It is easy to show that this matrix has rank 2, and hence that the subspace U is generated just by the first two matrices, which consequently form a basis for U. In fact, it follows easily that

$$\begin{pmatrix} -1 & 0 \\ 1 & 2 \\ 3 & 4 \end{pmatrix} = -1 \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} + 2 \cdot \begin{pmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}$$

$$\begin{pmatrix} -2 & -1 \\ 0 & 1 \\ 2 & 3 \end{pmatrix} = -2 \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} + 3 \cdot \begin{pmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}.$$

There are many options for extending this basis to be a basis for the whole space. One option is to choose the first two of the given matrices together with the last four (actually, any four would do) of the six linearly independent matrices

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

. Linear independence of these six matrices is established by computing

$$\begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} = 1 \neq 0.$$

Clearly the dimension is 6, so spanning is automatic, and hence we have a basis. □

How can we describe simple mappings analytically? For example, how can we describe a rotation, an axial symmetry, a mirror symmetry, a projection of a three-dimensional space onto a two-dimensional one in the plane or in the space? How can we describe the scaling of a diagram? What do they have in common? These all are linear mappings. This means that they preserve a certain structure of the space or a subspace. What structure? The structure of a vector space. Every point in the plane is described by two coordinates, every point in the 3-dimensional space is described by three coordinates. If we fix the origin, then it makes sense to say that a point is in some direction twice that far from the origin as some other point.

A mapping assigning the vector $v = a_1 v_1 + \cdots + a_n v_n$ to its coordinates in the basis $\underline{v}$ will be denoted by the same symbol $\underline{v} : V \to \mathbb{K}^n$. 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 $\mathbb{R}^2$:

Let V and W be vector spaces over the same field of scalars $\mathbb{K}$. The mapping $f : V \to W$ 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:

$$f : \mathbb{K}^n \to \mathbb{K}^m, \quad x \mapsto A \cdot x$$

with a fixed matrix A of the type m/n over $\mathbb{K}$.

The *image* of a linear mapping, $\operatorname{Im} f = f(V) \subset W$, 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 $\operatorname{Ker} f = f^{-1}(\{0\}) \subset V$ is a subspace, since the linear combination of zero images will always be a zero vector. The subspace $\operatorname{Ker} f$ is called the *kernel of the linear mapping f*.

A linear mapping which is a bijection is called an *isomorphism*.

Analogously to the abstract definition of vector spaces, it is again necessary to prove seemingly trivial claims that follow from the axioms:

We also know where arrive at if we translate or shift by some amount in a given direction and then by some other amount in another direction. These properties can be formalized – we speak of vectors in the plane or in space, and we consider their multiplication and addition. Linear mappings have the property that the image of a sum of vectors is a sum of the images of the vectors. The image of a multiple of a vector is the same multiple as the image of the vector. These properties are shared among the mappings stated at the beginning of this paragraph. Such a mapping is then uniquely determined by its behaviour on the vectors of a basis. (In the plane, a basis consists of two vectors not on the same line. In space a basis consists of three vectors not all in the same plane).

How can we write down some linear mapping f on a vector space V? For simplicity, we start with the plane $\mathbb{R}^2$. Assume that the image of the point (vector) $(1,0)$ is (a,b) and the image of the point (vector) $(0,1)$ is (c,d). This uniquely determines the image of an arbitrary point with coordinates (u,v): $f((u,v)) = f(u(1,0)+v(0,1)) = uf(1,0)+ vf(1,0) = (ua, ub) + (vc, vd) = (au + cv, bu + dv)$. This can be written down more efficiently as follows:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} au + cv \\ bu + dv \end{pmatrix}$$

A linear mapping is thus a mapping uniquely determined (in a fixed basis) by a matrix. Furthermore, when we have another linear mapping g given by the matrix $\begin{pmatrix} e & f \\ g & h \end{pmatrix}$, then we can easily compute (an interested reader can fill in the details by himself) that their composition $g \circ f$ is given by the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + fc & be + df \\ ag + ch & bg + dh \end{pmatrix}.$$

This leads us to the definition of matrix multiplication in exactly this way. That is, an application of a mapping on a vector is given by the matrix multiplication of the matrix of the mapping with the given vector, and that the mapping of a composition is given by the product of the corresponding matrices. This works analogously in the spaces of higher dimension. Further, this again shows what has already been proven in (2.1.5), namely, that matrix multiplication is associative but not commutative, just as with mapping composition. That is another motivation to study vector spaces.

Recall that already in the first chapter we worked with the matrices of some linear mappings in the plane $\mathbb{R}^2$, notably with the rotation around a point and with axial symmetry (see 1.5.8 and 1.5.9).

Proposition. *Let $f : V \to W$ be a linear mapping between two vector spaces over the same field of scalars $\mathbb{K}$. The following is true for all vectors $u, u_1, \ldots, u_k \in V$ and scalars $a_1, \ldots, a_k \in \mathbb{K}$*

(1) $f(0) = 0$,

(2) $f(-u) = -f(u)$,

(3) $f(a_1 \cdot u_1 + \cdots + a_k \cdot u_k) = a_1 \cdot f(u_1) + \cdots + a_k \cdot f(u_k)$,

(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 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 $f(a \cdot u + b \cdot v) = a \cdot f(u) + b \cdot f(v) \in W_1$. □

2.3.13. Proposition (Simple corollaries). *(1) The composition $g \circ f : V \to Z$ of two linear mappings $f : V \to W$ and $g : W \to Z$ is again a linear mapping.*

(2) The linear mapping $f : V \to W$ is an isomorphism if and only if $\operatorname{Im} f = W$ and $\operatorname{Ker} f = \{0\} \subset V$. The inverse mapping of an isomorphism is again an isomorphism.

(3) For any two subspaces V_1, $V_2 \subset V$ and linear mapping $f : V \to W$,

$$f(V_1 + V_2) = f(V_1) + f(V_2),$$
$$f(V_1 \cap V_2) \subset f(V_1) \cap f(V_2).$$

(4) The "coordinate assignment" mapping $\underline{u} : V \to \mathbb{K}^n$ given by an arbitrarily chosen basis $\underline{u} = (u_1, \ldots, u_n)$ 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.

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 $\operatorname{Im} f = W$. If $\operatorname{Ker} f = \{0\}$ 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

$$f(w) = au + bv = f(a \cdot f^{-1}(u) + b \cdot f^{-1}(v)).$$

We try now to write down matrices of linear mappings from $\mathbb{R}^3$ to $\mathbb{R}^3$. What does the matrix of a rotation in three dimensions look like? We begin with some special (easier for description) rotations about coordinate axes:

2.C.18. Matrix of rotation about coordinate axes in $\mathbb{R}^3$. We write down the matrices of rotations by the angle φ, about the (oriented) axes x, y and z in $\mathbb{R}^3$.

Solution. When rotating a particular point about the given axis (say x), the corresponding coordinate (x) does not change. The remaining two coordinates are then given by the rotation in the plane which we already know (a matrix of the type 2×2).

Thus we obtain the following matrices – rotation about the axis z:

$$\begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

rotation about the axis x:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}.$$

rotation about the axis y:

$$\begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix}$$

Note the sign of φ in the matrix for rotation about y. We want, as with any other rotation, the rotation about the y axis to be in the positive sense — that is, when we look in the opposite direction of the direction of the y axis, the world turns anti-clockwise. The signs in the matrices depend on the orientation of our coordinate system. Usually, in the 3-dimensional space the "dextrorotary coordinate system" is chosen: if we place our hand on the x axis such that the fingers point in the direction of the axis and such that we can rotate the x axis in the xy plane so that x coincides with the y axis and they point in the same direction, then the thumb should point in the direction of the z axis. In such a system, this is a rotation in the negative sense in the plane xz (that is, the axis z turns in the direction towards x). Think about the positive and negative sense of rotations by all three axes. The sign is also consistent with the cycle x to y to z to x to y etc.... or 1 to 2 to 3 to 1 to..... etc. □

Knowledge of matrices allows us to write the matrix of rotation about any oriented axis. Let us start with a specific example:

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 by sharp.

The remaining claims all follow immediately from the definition. □

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 $f : V \to W$. For every choice of basis $\underline{u} = (u_1, \ldots, u_n)$ on V, $\underline{v} = (v_1, \ldots, v_n)$ on W there are the following linear mappings as shown in the diagram:

$$
\begin{array}{ccc}
V & \xrightarrow{\;\;f\;\;} & W \\
{\scriptstyle \underline{u}}\downarrow{\scriptstyle\simeq} & & {\scriptstyle\simeq}\downarrow{\scriptstyle \underline{v}} \\
\mathbb{K}^n & \xdashrightarrow{f_{\underline{u},\underline{v}}} & \mathbb{K}^m
\end{array}
$$
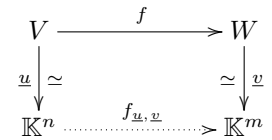
The bottom arrow $f_{\underline{u},\underline{v}}$ is defined by the remaining three, i.e. the composition of linear mappings

$$f_{\underline{u},\underline{v}} = \underline{v} \circ f \circ \underline{u}^{-1}.$$

<div style="text-align:center">MATRIX OF A LINEAR MAPPING</div>

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

$$f(u_1) = a_{11} \cdot v_1 + a_{21} \cdot v_2 + \cdots + a_{m1}v_m$$
$$f(u_2) = a_{12} \cdot v_1 + a_{22} \cdot v_2 + \cdots + a_{m2}v_m$$
$$\vdots$$
$$f(u_n) = a_{1n} \cdot v_1 + a_{2n} \cdot v_2 + \cdots + a_{mn}v_m,$$

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 $\underline{u}, \underline{v}$.

For a general vector $u = x_1 u_1 + \cdots + x_n u_n \in V$ we calculate (recall that vector addition is commutative and distributive with respect to scalar multiplication)

$$f(u) = x_1 f(u_1) + \cdots + x_n f(u_n)$$
$$= x_1(a_{11}v_1 + \cdots + a_{m1}v_m) + \cdots + x_n(a_{1n}v_1 + \cdots)$$
$$= (x_1 a_{11} + \cdots + x_n a_{1n})v_1 + \cdots + (x_1 a_{m1} + \cdots)v_m.$$

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 $\mathbb{K}^n \to \mathbb{K}^m$ and thus also a mapping

2.C.19. 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 $(1, 1, 0)$ under the standard basis $\mathbb{R}^3$.

Solution. The given rotation is easily obtained by composing these 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

$$\begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{6}}{4} \\ \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{6}}{4} \\ -\frac{\sqrt{6}}{4} & \frac{\sqrt{6}}{4} & \frac{1}{2} \end{pmatrix}$$

$\square$

$f : V \to W$. We have found the bijective correspondence between matrices of the fixed types (determined by dimensions of V and W) and linear mappings $V \to W$.

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 id_V.

$$\begin{array}{ccc} V & \xrightarrow{\mathrm{id}_V} & V \\ {\scriptstyle \underline{u}} \downarrow {\scriptstyle \simeq} & & {\scriptstyle \simeq} \downarrow {\scriptstyle \underline{v}} \\ \mathbb{K}^n & \xdashrightarrow{\ T=(\mathrm{id}_V)_{\underline{u},\underline{v}}\ } & \mathbb{K}^n \end{array}$$

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 $\underline{u}$

$$u = x_1 u_1 + \cdots + x_n u_n,$$

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:

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 $g : W \to Z$ and denote the corresponding matrix by $g_{\underline{v},\underline{w}}$.

2.C.20. Matrix of general rotation in $\mathbb{R}^3$. Derive the matrix of a general rotation in $\mathbb{R}^3$.

Solution. We can do the same things as in the previous example with general values. Consider an arbitrary unit vector (x, y, z). Rotation in the positive sense by the angle φ about this vector can be written down as a composition of the following rotations whose matrices we already know:

i) 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) goes over on the line with the directional vector $(0, y, z)$. The matrix of this rotation is
$$R_1 = \begin{pmatrix} x/\sqrt{1-z^2} & y/\sqrt{1-z^2} & 0 \\ -y/\sqrt{1-z^2} & x/\sqrt{1-z^2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

ii) 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 $(0, y, z)$ goes over on the line with the directional vector $(1, 0, 0)$. The matrix of this rotation is
$$R_2 = \begin{pmatrix} \sqrt{1-z^2} & 0 & z \\ 0 & 1 & 0 \\ -z & 0 & \sqrt{1-z^2} \end{pmatrix},$$

iii) rotation $\mathcal{R}_3$ in the positive sense about the x axis through the angle φ with the matrix
$$R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix},$$

iv) rotation $\mathcal{R}_2^{-1}$ with the matrix R_2^{-1},

v) rotation $\mathcal{R}_1^{-1}$ with the matrix R_1^{-1}.

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 $s = \sin\varphi$.

$\square$

We got familiar with matrices of linear maps now. But what happen with the matrix of a linear mapping, if we change the base of the vector space? (we can imagine it for example
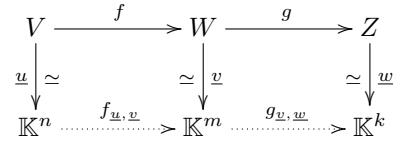
The composition $g \circ f$ on the upper row corresponds to the matrix of the mapping $\mathbb{K}^n \to \mathbb{K}^k$ on the bottom and we calculate directly (we write A for the matrix of f and B for the matrix of g in the chosen bases):
$$g_{\underline{v},\underline{w}} \circ f_{\underline{u},\underline{v}}(x) = \underline{w} \circ g \circ \underline{v}^{-1} \circ \underline{v} \circ f \circ \underline{u}^{-1}$$
$$= B \cdot (A \cdot x) = (B \cdot A) \cdot x = (g \circ f)_{\underline{u},\underline{w}}(x)$$
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:
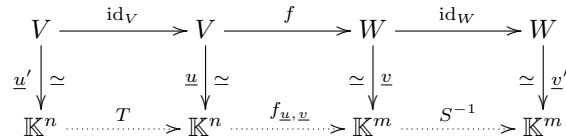
where T is the coordinate transition matrix from $\underline{u}'$ to $\underline{u}$ and S is the coordinate change matrix from $\underline{v}'$ to $\underline{v}$. If A is the original matrix of the mapping, then the matrix of the new mapping is given by $A' = S^{-1}AT$.

In the special case of a linear mapping $f : V \to V$, 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 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 $f : V \to \mathbb{K}$. 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 $\underline{v} = (v_1, \ldots, v_n)$, there are the linear forms $v_i^* : V \to \mathbb{K}$ 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^*, $\underline{v}^* = (v_1^*, \ldots, v_n^*)$, 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

as the change of coordinate system of the observer) We have to understand, what happens with the coordinates of vectors first. The key to all this is the transition matrix (see 2.3.15). We will further write $\underline{e}$ for the standard basis, that is vectors $((1, 0, 0), (0, 1, 0), (0, 0, 1))$ (these vectors could be any three linearly independent vectors in a vector space; with naming them as we did, we identified the vector space with $\mathbb{R}^3$)

2.C.21. A vector has coordinates $(1, 2, 3)$ in the standard basis $\underline{e}$. What are its coordinates in the basis $\underline{u} = ((1, 1, 0), (1, -1, 2), (3, 1, 5))$?

Solution. We write the transiton matrix T for $\underline{u}$ to the standard basis first. We just write coordinates of the vectors which form the basis $\underline{u}$ in the columns:

$$T = \begin{pmatrix} 1 & 1 & 3 \\ 1 & -1 & 1 \\ 3 & 1 & 5 \end{pmatrix}.$$

For expressing the sought coordinates we albeit need the transition matrix from the standard basis to $\underline{u}$. No problem, it is just $T^{(-1)}$. (see 2.3.15 if you have not done so yet). We already know how to compute inverse matrix (see 2.1.10).

$$T^{-1} = \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & 1 \\ -\frac{1}{2} & -1 & \frac{1}{2} \\ 1 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Finally the sought coordinates are

$$T^{-1}(1, 2, 3)^T = (\frac{1}{2}, -1, \frac{1}{2})^T.$$

$\square$

Similarly we work with the matrix of a linear mapping.

2.C.22. We are given a linear mapping $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ in the standard basis as the following matrix:

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 0 \end{pmatrix}.$$

Write down the matrix of this mapping in the basis $(f_1, f_2, f_3) = ((1, 1, 0), (-1, 1, 1), (2, 0, 1))$.

Solution. 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 writing down the coordinates of the vectors f_1, f_2, f_3 in the standard basis as the columns of the matrix T. Thus we have

$$T = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

it!) and if $\alpha \in V^*$ is an arbitrary form, then for every vector $u = x_1 v_1 + \cdots + x_n v_n$

$$\alpha(u) = x_1 \alpha(v_1) + \cdots + x_n \alpha(v_n)$$
$$= \alpha(v_1) v_1^*(u) + \cdots + \alpha(v_n) v_n^*(u)$$

and thus the linear form α 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 α with matrices of the type $1/n$, that is, with rows y. The components of these rows are coordinates of the general linear forms α 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 $\underline{v}$:

$$\alpha(u) = y \cdot x = y_1 x_1 + \cdots + y_n x_n.$$

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 70.

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, $i = 0, 1, 2, \ldots$.

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.

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 φ 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

The transition matrix for changing the basis from the standard basis to the basis $\underline{f}$ is then the inverse of T:

$$T^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{1}{2} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

The matrix of the mapping in the basis $\underline{f}$ is then given by (see 2.2.11)

$$T^{-1}AT = \begin{pmatrix} \frac{1}{4} & 2 & -\frac{3}{4} \\ \frac{5}{4} & 0 & \frac{7}{4} \\ \frac{3}{4} & -2 & \frac{9}{4} \end{pmatrix}.$$

□

2.C.23. Consider the vector space of polynomials of one variable of degree at most 2 with real coefficients. In this space, consider the basis 1, x, x^2. Write down the matrix of the derivative mapping in this basis and also in the basis $\underline{f} = (1 + x^2, x, x + x^2)$.

Solution. First we have to determine the matrix of the derivative mapping (let us denote the mapping as d, its matrix as D). We chose the basis $(1, x, x^2)$ as a standard basis $\underline{e}$, so we have coordinates $1 \sim (1,0,0)$, $x \sim (0,1,0)$ and $x^2 \sim (0,0,1)$. We look at the images of the basis vectors: $d(1) = 0 \sim (0,0,0)$, $d(x) = 1 \sim (1,0,0)$ and $d(x^2) = 2x \sim (0,2,0)$. Now we write the images as columns into the matrix D:

$$D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now we write the coordinates of the basis vectors of the basis $\underline{f}$ into the columns:

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

to get the transition matrix from $\underline{f}$ to $\underline{e}$. As in the previous example we get the matrix of d in the basis $\underline{f}$ as

$$T^{-1}DT = \begin{pmatrix} 0 & 1 & 1 \\ 2 & 1 & 3 \\ 0 & -1 & -1 \end{pmatrix},$$

where we had to compute

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

□

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 $\langle\,,\,\rangle : V \times V \to \mathbb{R}$ 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 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 $\underline{u} = (u_1, \ldots, u_n)$ of the space V and denote

$$s_{ij} = \langle u_i, u_j \rangle.$$

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*

$$\langle x, y \rangle = y^T \cdot x.$$

For each basis of the space V there is the symmetric matrix S such that the coordinate expression of the scalar product is

$$\langle x, y \rangle = y^T \cdot S \cdot x.$$

2.C.24. In the standard basis in $\mathbb{R}^3$, determine the matrix of the rotation through the angle $90°$ in the positive sense about the line (t, t, t), $t \in \mathbb{R}$, oriented in the direction of the vector $(1, 1, 1)$. Further, find the matrix of this rotation in the basis $\underline{g} = ((1, 1, 0), (1, 0, -1), (0, 1, 1))$.

Solution. We can easily determine the matrix of the given rotation in a suitable basis, that is, in a basis given by the directional vector of the line and by two mutually perpendicular vectors in the plane $x + y + z = 0$, that is, in the plane of vectors perpendicular to the vector $(1, 1, 1)$. We note that the matrix of the rotation in the positive sense through $90°$ in an orthonormal basis in $\mathbb{R}^2$ is $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. In the orthogonal basis with vectors of length k, l respectively, it is $\begin{pmatrix} 0 & -k/l \\ l/k & 0 \end{pmatrix}$. If we choose perpendicular vectors $(1, -1, 0)$ and $(1, 1, -2)$ in the plane $x + y + z = 0$ with lengths $\sqrt{2}$ and $\sqrt{6}$, then in the basis $\underline{f} = ((1, 1, 1), (1, -1, 0), (1, 1, -2))$ the rotation we are looking for has matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -\sqrt{3} \\ 0 & 1/\sqrt{3} & 0 \end{pmatrix}$. In order to obtain the matrix of the rotation in the standard basis, it is enough to change the basis. The transition matrix T for changing the basis from the basis $\underline{f}$ to the standard basis is obtained by writing the coordinates (under the standard basis) of the vectors of the basis $\underline{f}$ as the columns of the matrix T: $T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & -2 \end{pmatrix}$. Finally, for the desired matrix R, we have

$$R = T \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -\sqrt{3} \\ 0 & 1/\sqrt{3} & 0 \end{pmatrix} \cdot T^{-1}$$

$$= \begin{pmatrix} 1/3 & 1/3 - \sqrt{3}/3 & 1/3 + \sqrt{3}/3 \\ 1/3 + \sqrt{3}/3 & 1/3 & 1/3 - \sqrt{3}/3 \\ 1/3 - \sqrt{3}/3 & 1/3 + \sqrt{3}/3 & 1/3 \end{pmatrix}$$

This result can be checked by substituting into the matrix of general rotation (2.C.20). By normalizing the vector $(1, 1, 1)$ we obtain the vector $(x, y, z) = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, $\cos(\varphi) = 0$, $\sin(\varphi) = 1$. $\square$

2.C.25. Matrix of general rotation revisited. We derive the matrix of (general) rotation from (2.C.20) through the angle φ in the positive sense about the unit vector (x, y, z) in a different way, analogously to the previous exercise. In the basis

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 $(u_1, \ldots, u_k)$ 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 $f : V \to V$ 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 $v \in V$,

$$v = f(v) + (v - f(v)) \in \mathrm{Im}(f) + \mathrm{Ker}(f) = V$$

and if $v \in \mathrm{Im}(f)$ 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 $W = \mathrm{Im}(f)$ along the subspace $U = \mathrm{Ker}(f)$. 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.

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. 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. Immediately we have a scalar product. Call it orthonormal. In this basis the scalar products of vectors are computed as in the formula in the Theorem 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*.

$\underline{f} = ((x, y, z), (-y, x, 0), (zx, zy, z^2 - 1))$, that is, in the orthogonal basis composed of the directional vector of the axis of rotation and of two mutually perpendicular vectors with sizes $\sqrt{1 - z^2}$ lying in a plane perpendicular to the axis of rotation, the matrix corresponding to the rotation is $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix}$. The matrix for changing the basis from $\underline{f}$ to the standard basis is then $T = \begin{pmatrix} x & -y & zx \\ y & x & zy \\ z & 0 & z^2 - 1 \end{pmatrix}$ with the inverse matrix

$$T^{-1} = \begin{pmatrix} x & y & z \\ -\frac{y}{1-z^2} & \frac{x}{1-z^2} & 0 \\ \frac{zx}{1-z^2} & \frac{zy}{1-z^2} & -1 \end{pmatrix}.$$

Finally, for the matrix R of the rotation we obtain

$$R = T \cdot A \cdot T^{-1}$$

$$= \begin{pmatrix} 1 - t + tx^2 & txy - zs & txz + ys \\ yxt + zs & 1 - t + ty^2 & tyz - xs \\ zxt - ys & tzy + xs & 1 - t + tz^2 \end{pmatrix},$$

where again $t = 1 - \cos\varphi$ and $s = \sin\varphi$, and we get the same matrix as before.

When multiplying and simplifying, we must repeatedly use the assumption $x^2 + y^2 + z^2 = 1$.

Through a more detailed analysis of properties of various types of linear mapping we now obtain a deeper understanding of tools we are given by vector spaces for linear modeling of processes and systems.

2.C.26. Consider complex numbers as a real vector space and choose 1 and i for its basis. Determine in this basis the matrix of the following linear mappings:

  a) conjugation,

  b) multiplication by the number $(2 + i)$.

Determine the matrix of these mappings in the basis $\underline{f} = ((1 - i), (1 + i))$.

Solution. In order to determine the matrix of a linear mapping in some basis, it is enough to determine the images of the basis vectors.

  a) For conjugation we have $1 \mapsto 1$, $i \mapsto -i$, written in the coordinates $(1, 0) \mapsto (1, 0)$ and $(0, 1) \mapsto (0, -1)$. By writing the images into the columns we obtain the matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, In the basis $\underline{f}$ the conjugation interchanges basis vectors, that is, $(1, 0) \mapsto (0, 1)$ and $(0, 1) \mapsto (1, 0)$ and the matrix of conjugation under this basis is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

The point of this procedure is to transform a given sequence of independent generators $v_1, \ldots, v_k$ of a finite dimensional space V into an orthogonal set of independent generators of V.

GRAMM-SCHMIDT ORTHOGONALIZATION

Proposition. *Let $(u_1, \ldots, u_k)$ be a linearly independent k-tuple of vectors of a space V with scalar product. Then there exists an orthogonal system of vectors $(v_1, \ldots, v_k)$ such that $v_i \in \mathrm{span}\{u_1, \ldots, u_i\}$, and $\mathrm{span}\{u_1, \ldots, u_i\} = \mathrm{span}\{v_1, \ldots, v_i\}$, for all $i = 1, \ldots, k$. 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 $v_1, \ldots, v_\ell$ with the required properties and if $\ell < k$, we choose $v_{\ell+1} = u_{\ell+1} + a_1 v_1 + \cdots + a_\ell v_\ell$, where $a_i = -\frac{\langle u_{\ell+1}, v_i \rangle}{\|v_i\|^2}$.*

PROOF. We begin with the first (nonzero) vector v_1 and calculate the orthogonal projection v_2 to

$$\mathrm{span}\{v_1\}^{\perp} \subset \mathrm{span}\{v_1, v_2\}.$$

The result is nonzero if and only if v_2 is independent of v_1. All other steps are similar:

In step $\ell, \ell > 1$ we seek the vector $v_{\ell+1} = u_{\ell+1} + a_1 v_1 + \cdots + a_\ell v_\ell$ satisfying $\langle v_{\ell+1}, v_i \rangle = 0$ for all $i = 1, \ldots, \ell$. This implies

$$0 = \langle u_{\ell+1} + a_1 v_1 + \cdots + a_\ell v_\ell, v_i \rangle = \langle u_{\ell+1}, v_i \rangle + a_i \langle v_i, v_i \rangle$$

and we can see that the vectors with the desired properties are determined uniquely up to a scalar multiple. $\square$

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 $(e_1, \ldots, e_n)$ for a space V. Then every vector $v = x_1 e_1 + \cdots + x_n e_n$ satisfies

$$\langle e_i, v \rangle = \langle e_i, x_1 e_1 + \cdots + x_n e_n \rangle = x_i$$

and so we can always express

$$(1) \qquad v = \langle e_1, v \rangle e_1 + \cdots + \langle e_n, v \rangle e_n.$$

If we are given a subspace $W \subset V$ and its orthonormal basis $(e_1, \ldots, e_k)$, then we can extend it to an orthonormal basis $(e_1, \ldots, e_n)$ for V. Orthogonal projection of a general vector $v \in V$ to W is then given by the expression

$$v \mapsto \langle e_1, v \rangle e_1 + \cdots + \langle e_k, v \rangle e_k.$$

b) For the basis $(1, i)$ 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 $(1, i)$ is:
$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}.$$

We determine the matrix in the basis $\underline{f}$. 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 $\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$.

Think about the following: why is the matrix of multiplication by $2 + i$ the same in both bases? Would the two matrices in these bases be the same for multiplication by any complex number? □

2.C.27. Determine the matrix A which, under the standard basis of the space $\mathbb{R}^3$, gives the orthogonal projection on the vector subspace generated by the vectors $u_1 = (-1, 1, 0)$ and $u_2 = (-1, 0, 1)$.

Solution. Note first that the given subspace is a plane containing the origin with normal vector $u_3 = (1, 1, 1)$. The ordered triple $(1, 1, 1)$ is clearly a solution to the system
$$\begin{array}{rcrcrcl} -x_1 & + & x_2 & & & = & 0, \\ -x_1 & & & + & x_3 & = & 0, \end{array}$$
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 on the zero vector. In the basis composed of u_1, u_2, u_3 (in this order) is thus the matrix of this projection
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Using the the transition matrix for changing the basis
$$T = \begin{pmatrix} -1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$
from the basis (u_1, u_2, u_3) to the standard basis, and from the standard basis to the basis (u_1, u_2, u_3) we obtain
$$A = \begin{pmatrix} -1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$
□

In particular, we need only consider an orthonormal basis of the subspace W in order to write the orthogonal projection to W explicitly.

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 $g = \mathrm{id}_V - f$. 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 $\mathbb{R}^n$.

The constant coefficient is the determinant $|A|$. We shall see later that this coefficient describes how much the linear mapping scales the volumes.

We shall return to the questions of the length of a vector and to projections in the following chapter in a more general context.

2.3.21. 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 φ 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.

2.3.22. 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.

D. Properties of linear maps

2.D.1. Write down the matrix of the mapping of orthogonal projection on the plane passing through the origin and perpendicular to the vector $(1, 1, 1)$.

Solution. The image of an arbitrary point (vector) $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ under the considered mapping can be obtained by subtracting from the given vector its orthogonal projection onto the direction normal to the considered plane, that is, onto the direction $(1, 1, 1)$. This projection $\mathbf{p}$ is given by (see 1) as

$$\frac{\langle \mathbf{x}, (1, 1, 1) \rangle}{|(1, 1, 1)|^2}$$
$$= \left(\frac{x_1 + x_2 + x_3}{3}, \frac{x_1 + x_2 + x_3}{3}, \frac{x_1 + x_2 + x_3}{3} \right).$$

The resulting mapping is thus

$$\mathbf{x} - \mathbf{p}$$
$$= \left(\frac{2x_1}{3} - \frac{x_2 + x_3}{3}, \frac{2x_2}{3} - \frac{x_1 + x_3}{3}, \frac{2x_3}{3} - \frac{x_1 + x_2}{3} \right) =$$
$$= \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

We have (correctly) obtained the same matrix as in the exercise 2.C.27. □

2.D.2. In $\mathbb{R}^3$ write down the matrix of the mirror symmetry with respect to the plane containing the origin and $(1, 1, 1)$ being its normal vector.

Solution. As in 2.D.1 we get the image of an arbitrary vector $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ with the help of the orthogonal projection onto the direction $(1, 1, 1)$. Unlike in the previous example, we need to subtract the projection twice (see image). Thus we get the matrix:

$$x - 2p =$$
$$\left(\frac{x_1}{3} - \frac{2(x_2 + x_3)}{3}, \frac{x_2}{3} - \frac{2(x_1 + x_3)}{3}, \frac{x_3}{3} - \frac{2(x_1 + x_2)}{3} \right)$$
$$= \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Second solution. The normed normal vector of the mirror plane is $n = \frac{1}{\sqrt{3}}(1, 1, 1)$. We can express the mirror image of v under the mirror symmetry Z as follows: $Z(v) = v - 2\langle v, n \rangle n = v - 2n \cdot (n^T \cdot v) = v - 2(n \cdot n^T) \cdot v = ((E - 2n \cdot n^T)v$ (where we have used $\langle v, n \rangle = v \cdot n^T$ for the standard scalar product and the associativity of the matrix

Most often we will meet *bilinear forms*, that is, the case $\alpha : V \times V \to \mathbb{K}$, where for any four vectors u, v, w, z and scalars a, b, c and d we have

$$\alpha(au + bv, cw + dz) = ac\,\alpha(u, w) + ad\,\alpha(u, z)$$
$$+ bc\,\alpha(v, w) + bd\,\alpha(v, z).$$

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 α, 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 α scalars $a_{ij} = \alpha(u_i, u_j)$ then we obtain for vectors v, w with coordinates x and y (as columns of coordinates)

$$\alpha(v, w) = \sum_{i,j=1}^{n} a_{ij} x_i y_j = x^T \cdot A \cdot y,$$

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 α on a vector space V defines a mapping $V \to V^*$, $v \mapsto \alpha(v, \ )$. That is, by placing a fixed vector in the first 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

$$x \mapsto (y \mapsto x^T \cdot A \cdot y).$$

All this is a matter of convention. Also we may fix the second 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.

multiplication). We get the same matrix:

$$E - 2n \cdot n^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{2}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$= \frac{1}{3} \begin{pmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{pmatrix}.$$

$\square$

2.D.3. Consider $\mathbb{R}^3$, with the standard coordinate system. In the plane $z = 0$ there is a mirror and at the point $[4, 3, 5]$ there is a candle. The observer at the point $[1, 2, 3]$ is not aware of the mirror, but sees in it the reflection of the candle. Where does he think the candle is?

Solution. Independently of our position, we see the mirror image of the scene in the mirror (that is why it is called a mirror image). The mirror image is given by reflecting the scene (space) by the plane of the mirror, the plane $z = 0$. The reflection with respect to this plane changes the sign of the z-coordinate. That is we can see the candle at the point $[4, 3, -5]$. $\square$

By using the inner product we can determine the (angular) deflection of the vectors:

2.D.4. Determine the deflection of the roots of the polynomial $x^2 - i$ considered as vectors in the complex plane.

Solution. The roots of the given polynomial are square roots of i. The arguments of the square roots of any complex numbers differ according to the de Moivre theorem by π. Their deflection is thus always π. $\square$

2.D.5. Determine the cosine of the deflection of the lines p, q in $\mathbb{R}^3$ given by the equations

$$\begin{aligned} p \quad : \quad & -2x + y + z = 1 \\ & x + 3y - 4z = 5 \\ q \quad : \quad & x - y = -2 \\ & z = 6 \end{aligned}$$

$\bigcirc$

2.D.6. Using the Gram-Schmidt orthogonalisation, obtain the orthogonal basis of the subspace

$$U = \left\{ (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4; \ x_1 + x_2 + x_3 + x_4 = 0 \right\}$$

of the space $\mathbb{R}^4$.

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 $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, C = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

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 $f : \mathbb{R}^2 \to \mathbb{R}^2$ 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 $g : \mathbb{C}^2 \to \mathbb{C}^2$ 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

$$g(u) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i \cdot u,$$

$$g(v) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} -i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -i \end{pmatrix} = -i \cdot v.$$

Solution. The set of solutions of the given homogeneous linear equation is clearly a vector space with the basis

$$u_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

shall be denoted Denote by v_1, v_2, v_3, vectors of the orthogonal basis obtained using the Gram-Schmidt orthogonalisation process.

First set $v_1 = u_1$. Then let

$$v_2 = u_2 - \frac{u_2^T \cdot v_1}{||v_1||^2} v_1 = u_2 - \frac{1}{2} v_1 = \left(-\frac{1}{2}, -\frac{1}{2}, 1, 0\right)^T,$$

that is, choose a multiple $v_2 = (-1, -1, 2, 0)^T$. Then let

$$v_3 = u_3 - \frac{u_3^T \cdot v_1}{||v_1||^2} v_1 - \frac{u_3^T \cdot v_2}{||v_2||^2} v_2 = u_3 - \frac{1}{2} v_1 - \frac{1}{6} v_2 =$$

$$= \left(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 1\right)^T.$$

Altogether we have

$$v_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 3 \end{pmatrix}.$$

Due to the simplicity of the exercise we can immediately give an orthogonal basis of the vectors

$$(1, -1, 0, 0)^T, \quad (0, 0, 1, -1)^T, \quad (1, 1, -1, -1)^T$$

or

$$(-1, 1, 1, -1)^T, \quad (1, -1, 1, -1)^T, \quad (-1, -1, 1, 1)^T.$$

$\square$

2.D.7. Write down a basis of the real vector space of the matrices 3×3 over $\mathbb{R}$ with zero trace. (The trace of a matrix is the sum of the elements on the diagonal). Write the coordinates of the matrix

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix}$$

in this basis.

2.D.8. Find the orthogonal complement $U^\perp$ of the subspace

$$U = \{(x_1, x_2, x_3, x_4); \, x_1 = x_3, x_2 = x_3 + 6x_4\} \subset \mathbb{R}^4.$$

Solution. The orthogonal complement $U^\perp$ consists of just those vectors that are perpendicular to every solution of the system

$$\begin{array}{ccccccc} x_1 & & - & x_3 & & = & 0, \\ & x_2 & - & x_3 & - & 6x_4 & = & 0. \end{array}$$

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 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 $f : V \to V$ 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)

$$(1) \qquad \det(A - a \cdot E) = 0.$$

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 $\mathbb{K}$.

A vector is a solution of this system if and only if it is perpendicular to both vectors $(1, 0, -1, 0)$, $(0, 1, -1, -6)$. Thus we have

$$U^\perp = \{a \cdot (1, 0, -1, 0) + b \cdot (0, 1, -1, -6);\ a, b \in \mathbb{R}\}.$$

$\square$

2.D.9. Find an orthonormal basis of the subspace $V \subset \mathbb{R}$, where $V = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_1 + 2x_2 + x_3 = 0\}$.

Solution. The fourth coordinate does not appear in the restriction for the subspace, thus it seems reasonable to select $(0, 0, 0, 1)$ as one of the vectors of the orthonormal basis and reduce the problem into the subspace $\mathbb{R}^3$. If we set the second coordinate equal to zero, then in the investigated space there are vectors with reverse first and third coordinate, notably, the unit vector $(\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}, 0)$. This vector is perpendicular to any vector which has first coordinate equal to the third coordinate. In order to get into the investigated subspace, we choose the second coordinate equal to the negative of the sum of the first and the third coordinate, and then normalise. Thus we choose the vector $(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, 0)$ and we are finished. $\square$

2.D.10. Find the eigenvalues and the associated subspaces
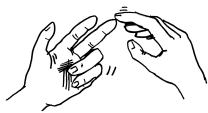
of eigenvectors of the matrix

$$A = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 3 & 0 \\ 2 & -2 & 2 \end{pmatrix}.$$

Solution. First we find the characteristic polynomial of the matrix:

$$\begin{vmatrix} -1-\lambda & 1 & 0 \\ -1 & 3-\lambda & 0 \\ 2 & -2 & 2-\lambda \end{vmatrix} = \lambda^3 - 4\lambda^3 + 2\lambda + 4.$$

This polynomial has roots $2, 1+\sqrt{3}, 1-\sqrt{3}$, which are then the eigenvalues of the matrix. Their algebraic multiplicity is one (they are simple roots of the polynomial), thus each has associated only one (up to a non-zero multiple) eigenvector. Otherwise stated, the geometric multiplicity of the eigenvalue is one, see 3.4.10).

We determine the eigenvector associated with the eigenvalue 2. It is a solution of the homogeneous linear system with the matrix $A - 2E$:

$$\begin{aligned} -3x_1 + x_2 &= 0 \\ -1x_1 + x_2 &= 0 \\ 2x_1 - 2x_2 &= 0. \end{aligned}$$

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 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 λ, 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 λ, together with the zero vector, form a nontrivial vector subspace $V_\lambda \subset V$. We call it the *eigenspace associated with λ*. For instance, if $\lambda = 0$ is an eigenvalue, the kernel $\operatorname{Ker} f$ 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:

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 $f : V \to V$ 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 $f : V \to V$ is independent of the choice

of the basis of V, the coefficients of individual powers of the variable λ 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

$$|A - \lambda \cdot E| = (-1)^n \lambda^n + (-1)^{n-1}(a_{11} + \cdots + a_{nn})\lambda^{n-1} + \cdots + |A|\lambda^0.$$

The coefficient at the highest power says whether the dimension of the space V is even or odd.

The system has solution $x_1 = x_2 = 0$, $x_3 \in \mathbb{R}$ arbitrary. So the eigenvector associated with the value 2 is then the vector $(0, 0, 1)$ (or any multiple of it).

Similarly we determine the remaining two eigenvectors – as solutions of the system $[A - (1 + \sqrt{3})E]\mathbf{x} = 0$. The solution of the system

$$
\begin{aligned}
(-2 - \sqrt{3})x_1 + x_2 &= 0 \\
-1x_1 + (2 - \sqrt{3})x_2 &= 0 \\
2x_1 - 2x_2 + (1 - \sqrt{3})x_3 &= 0
\end{aligned}
$$

is the space $\left\{ \left(2 - \sqrt{3}, 1, 2 \right) t, \ t \in \mathbb{R} \right\}$.

That is the space of eigenvectors associated with the eigenvalue $1 + \sqrt{3}$.

Similarly we obtain that the space of eigenvectors associated with the eigenvalue $1 - \sqrt{3}$ is $\left\{ \left(2 + \sqrt{3}, 1, -2 \right) t, \ t \in \mathbb{R} \right\}$. $\quad\square$

2.D.11. Determine the eigenvalues and eigenvectors of the matrix

$$
A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}.
$$

Describe the geometric interpretation of this mapping and write down its matrix in the basis:

$$
\begin{aligned}
e_1 &= (1, -1, 1) \\
e_2 &= (1, 2, 0) \\
e_3 &= (0, 1, 1)
\end{aligned}
$$

Solution. The characteristic polynomial of the matrix A is

$$
\begin{vmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 2 & 1-\lambda \end{vmatrix} = -\lambda^3 + 4\lambda^2 - 2\lambda = -\lambda(\lambda^2 - 4\lambda + 2).
$$

The roots of this polynomial are the eigenvalues, thus the eigenvalues are $0, 2 + \sqrt{2}, 2 - \sqrt{2}$. Thus eigenvalues are $0, 2 + \sqrt{2}, 2 - \sqrt{2}$. We compute the eigenvectors associated with the particular eigenvalues:

- 0: We solve the system

$$
\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0
$$

Its solutions form a one-dimensional vector space of eigenvectors: $\mathrm{span}\{(1, -1, 1)\}$.

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 $\mathrm{Tr}\, A$. 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 are just the trace. We shall see relation to matrix exponential later in Chapter 8.

The coefficient at λ^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 $f : V \to V$ associated to different eigenvalues are linearly independent.*

Proof. Let $a_1, \ldots, a_k$ be distinct eigenvalues of the mapping f and $u_1, \ldots, u_k$ eigenvectors with these eigenvalues. The proof is by induction on the number of linearly independent vectors among the chosen ones.

Assume that $u_1, \ldots, u_\ell$ are linearly independent and $u_{l+1} = \sum_i c_i u_i$ is their linear combination. We can choose $\ell = 1$, because the eigenvectors are nonzero. But then $f(u_{\ell+1}) = a_{l+1} \cdot u_{l+1} = \sum_{i=1}^{l} a_{l+1} \cdot c_i \cdot u_i$, that is,

$$
f(u_{l+1}) = \sum_{i=1}^{l} a_{l+1} \cdot c_i \cdot u_i = \sum_{i=1}^{l} c_i \cdot f(u_i) = \sum_{i=1}^{l} c_i \cdot a_i \cdot u_i.
$$

By subtracting the second and the fourth expression in the equalities we obtain $0 = \sum_{i=1}^{l} (a_{l+1} - a_i) \cdot c_i \cdot u_i$. 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 $u_1, \ldots, u_\ell$, therefore also the vector u_{l+1} must be linearly independent of the others. $\quad\square$

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 λ_i, we obtain the entire V as a direct sum of one-dimensional eigenspaces V_{λ_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 λ_i.

Furthermore, this decomposition can be easily calculated:

- $2 + \sqrt{2}$: We solve the system

$$\begin{pmatrix} -(1+\sqrt{2}) & 1 & 0 \\ 1 & -\sqrt{2} & 1 \\ 1 & 2 & -(1+\sqrt{2}) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0.$$

The solutions form a one-dimensional space $\mathrm{span}\{(1, 1+\sqrt{2}, 1+\sqrt{2})\}$.

- $2 - \sqrt{2}$: We solve the system

$$\begin{pmatrix} (\sqrt{2}-1) & 1 & 0 \\ 1 & \sqrt{2} & 1 \\ 1 & 2 & (\sqrt{2}-1) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0.$$

Its solutions form a space of eigenvectors $\mathrm{span}\{(1, 1-\sqrt{2}, 1-\sqrt{2})\}$.

Hence the given matrix has eigenvalues 0, $2 + \sqrt{2}$ and $2 - \sqrt{2}$, with the associated one-dimensional spaces of eigenvectors $\mathrm{span}\{(1, -1, 1)\}$, $\mathrm{span}\{(1, 1 + \sqrt{2}, 1 + \sqrt{2})\}$ and $\mathrm{span}\{(1, 1 - \sqrt{2}, 1 - \sqrt{2})\}$ respectively.

The mapping can thus be interpreted as a projection along the vector $(1, -1, 1)$ into the plane given by the vectors $(1, 1+\sqrt{2}, 1+\sqrt{2})$ and $(1, 1-\sqrt{2}, 1-\sqrt{2})$ composed with the linear mapping given by "stretching" by the factor corresponding to the eigenvalues in the directions of the associated eigenvectors.

Now we express it in the given basis. For this we need the matrix T for changing the basis from the standard basis to the new basis. This can be obtained by writing the coordinates of the vectors of the original basis under the new basis into the columns of the matrix T. But we shall do it in a different way – we obtain first the matrix for changing the basis from the new one to the original one, that is, the matrix T^{-1}. We just write the coordinates of the vectors of the new basis into the columns:

$$T^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Then

$$T = {T^{-1}}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ -2 & 1 & 3 \end{pmatrix},$$

and for the matrix B of a mapping under new basis we have (see 2.3.16)

$$B = TAT^{-1} = \begin{pmatrix} 0 & 5 & 2 \\ 0 & -2 & -1 \\ 0 & 14 & 6 \end{pmatrix}.$$

$\square$

You can find more exercises on computing with eigenvalues and eigenvectors on the page 133.

Corollary. *If there exist n mutually distinct roots λ_i of the characteristic polynomial of the mapping $f : V \to V$ 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 $f : V \to V$ generates a subspace $\mathrm{span}\{v\} \subset V$, 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 be a basis $(u_1, \dots, u_k, u_{k+1}, \dots, u_n)$ for the whole space V. For every such basis, the mapping will have a matrix A of the form

$$(1) \qquad A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$$

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 $W = \mathrm{span}\{u_1, \dots, u_k\}$ is invariant under the mapping f.

By the same arguments, the mapping with the matrix A as in (1) leaves the subspace $\mathrm{span}\{u_{k+1}, \dots, u_n\}$ 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

$$f : \mathbb{R}^3 \to \mathbb{R}^3, \ A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

In the case of a 3×3 matrix, you can use this special formula to find its characteristic polynomial:

2.D.12. For any $n \times n$ matrix A its characteristic polynomial $|A - \lambda E|$ is of degree n, that is, it is of the form

$$|A - \lambda E| = c_n \lambda^n + c_{n-1} \lambda^{n-1} + \cdots + c_1 \lambda + c_0, \quad c_n \neq 0,$$

while we have

$$c_n = (-1)^n, \quad c_{n-1} = (-1)^{n-1} \operatorname{tr} A, \quad c_0 = |A|.$$

If the matrix A is three-dimensional, we obtain

$$|A - \lambda E| = -\lambda^3 + (\operatorname{tr} A)\lambda^2 + c_1 \lambda + |A|.$$

By choosing $\lambda = 1$ we obtain

$$|A - E| = -1 + \operatorname{tr} A + c_1 + |A|.$$

From there we obtain

$$|A - \lambda E| =$$
$$-\lambda^3 + (\operatorname{tr} A)\lambda^2 + (|A - E| + 1 - \operatorname{tr} A - |A|)\lambda + |A|.$$

Use this expression for determining the characteristic polynomial and the eigenvalues of the matrix

$$A = \begin{pmatrix} 32 & -67 & 47 \\ 7 & -14 & 13 \\ -7 & 15 & -6 \end{pmatrix}.$$

○

2.D.13. Find the orthonormal complement of the vectorspace spaned by the vectors $(2, 1, 3)$, $(3, 16, 7)$, $(3, 5, 4)$, $(-7, 7, -10)$.

Solution. In fact the task consists of solving the system 2.A.3, which we have done already. □

2.D.14. Pauli matrices. In physics, the state of a particle with spin $\frac{1}{2}$ is described with Pauli matrices. They are the 2×2 matrices over complex numbers:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

For square matrices we define their *commutator* (denoted by square brackets) as $[\sigma_1, \sigma_2] := \sigma_1 \sigma_2 - \sigma_2 \sigma_1$

Show that $[\sigma_1, \sigma_2] = 2i\sigma_3$ and similarly $[\sigma_1, \sigma_3] = 2i\sigma_2$ and $[\sigma_2, \sigma_3] = 2i\sigma_1$. Furthermore, show that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$ and that the eigenvalues of the matrices σ_1, σ_2, σ_3 are ± 1.

Show that for matrices describing the state of the particle with spin 1, namely

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & -i & 0 \\ i & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

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 $f : \mathbb{R}_2[x] \to \mathbb{R}_2[x]$ 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 is no basis for which the matrix would be diagonal. There is the "chain" of vectors mapping four independent generators as follows: $\frac{1}{2}x^2 \mapsto x \mapsto 1 \mapsto 0$ builds a sequence of subspaces without invariant complements.

, the commuting relations are the same as in the case of Pauli matrices.

Equivalently it can be shown that under the notation $1 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, I := i\sigma_3, J := i\sigma_2, K := i\sigma_1$ forms the vector space with basis $(1, I, J, K)$ of an algebra of quaternions (the algebra is a vector space with binary bilinear operation of multiplication, in this case the multiplication is given by matrix multiplication). In order for the vector space to be an algebra of quaternions it is necessary and sufficient to show the following properties: $I^2 = J^2 = K^2 = -1$ and $IJ = -JI = K, JK = -KJ = I$ and $KI = -IK = J$.

2.D.15. Can the matrix

$$B = \begin{pmatrix} 5 & 6 \\ 6 & 5 \end{pmatrix}$$

be expressed in the form of the product $B = P^{-1} \cdot D \cdot P$ for some diagonal matrix D and invertible matrix P? If possible, give an example of such matrices D, P, and find out how many such pairs there are.

Solution. The matrix B has two distinct eigenvalues, and thus such an expression exists. For instance it holds that

$$\begin{pmatrix} 5 & 6 \\ 6 & 5 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sqrt{2} & -\sqrt{2} \\ \sqrt{2} & \sqrt{2} \end{pmatrix} \cdot \begin{pmatrix} 11 & 0 \\ 0 & -1 \end{pmatrix} \cdot \frac{1}{2} \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & \sqrt{2} \end{pmatrix}.$$

There exist exactly two diagonal matrices D:

$$\begin{pmatrix} 11 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \begin{pmatrix} -1 & 0 \\ 0 & 11 \end{pmatrix},$$

but the columns of the matrix P^{-1} can be substituted with their arbitrary non-zero scalar multiples, thus there are infinitely many pairs D, P. □

As we have already seen in 2.D.11, based on the eigenvalues and eigenvectors of the given 3×3 matrix, we can often interpret geometrically the mapping it induces in $\mathbb{R}^3$. In particular, we notice that can do so in the following situations: If the matrix has 0 as eigenvalue and 1 as an eigenvalue with geometric multiplicity 2, then it is a projection in the direction of the eigenvector associated with the eigenvalue 0 on the plane given by the eigenspace of the eigenvalue 1. If the eigenvector associated with 0 is perpendicular to that plane, then the mapping is an orthogonal projection.

If the matrix has eigenvalue -1 with the eigenvector perpendicular to the plane of the eigenvectors associated with the eigenvalue 1, then it is a mirror symmetry through the plane of the eigenvectors associated with 1.

2.4.6. Orthogonal mappings. We consider the special case of the mapping $f : V \to W$ between spaces with scalar products, which preserve lengths for all vectors $u \in V$.

A linear mapping $f : V \to W$ 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 $\mathbb{R}^2 \to \mathbb{R}^2$ 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 $f : V \to W$ 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 $f : V \to \text{Im } f$ 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:

Theorem. *Let V be a real vector space with scalar product and let $f : V \to V$ be a linear mapping. Then f is orthogonal if and only if in some orthogonal basis (and then consequently in all of them) its matrix A satisfies $A^T = A^{-1}$.*

If the matrix has eigenvalue 1 with an eigenvector perpendicular to plane of the eigenvectors associated with the eigenvalue -1, then it is an axial symmetry (in space) through the axis given by the eigenvector associated with 1.

2.D.16. Determine what linear mapping $\mathbb{R}^3 \to \mathbb{R}^3$ is given by the matrix

$$\begin{pmatrix} -\frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{4}{3} & -\frac{7}{3} & -\frac{8}{3} \\ -1 & 1 & 1 \end{pmatrix}$$

Solution. The matrix has a double eigenvalue -1, its associated eigenspace is $\mathrm{span}\{(2, 0, 1), (1, 1, 0)\}$. Further, the matrix has 0 as the eigenvalue, with eigenvector $(1, 4, -3)$. The mapping given by this matrix under the standard basis is then an axial symmetry through the line given by the last vector composed with the projection on the plane perpendicular to the last vector, that is, given by the equation $x + 4y - 3z = 0$.
□

2.D.17. The theorem 2.4.7 gives us tools for recognising a matrix of a rotation in $\mathbb{R}^3$. It is orthogonal (rows orthogonal to each other equivalently the same for the columns). It has three distinct eigenvalues with absolute value 1. One of them is the number 1 (its associated eigenvector is the axis of the rotation). The argument of the remaining two, which are necessarily complex conjugates, gives the angle of the rotation in the positive sense in the plane given by the basis $u_\lambda + \overline{u_\lambda}$, $i(u_\lambda - \overline{u_\lambda})$.

2.D.18. Determine what linear mapping is given by the matrix

$$\begin{pmatrix} \frac{3}{5} & \frac{16}{25} & \frac{-12}{25} \\ \frac{-16}{25} & \frac{93}{125} & \frac{24}{125} \\ \frac{12}{25} & \frac{24}{125} & \frac{107}{125} \end{pmatrix}.$$

Solution. First we notice, that the matrix is orthogonal (rows are mutually orhogonal, and equivalently the same with columns). The matrix has the following eigenvalues and corresponding eigenvectors: $1, v_1 = (0, 1, \frac{4}{3})$; $\frac{3}{5} + \frac{4}{5}i, v_2 = (1, \frac{4}{5}i, -\frac{3}{5}i)$; $\frac{3}{5} - \frac{4}{5}i, v_3 = (1, -\frac{4}{5}i, \frac{3}{5}i)$. All three eigenvalues have absolute value one, which together with the observation of orthogonality tells us that the matrix is a matrix of rotation. Its axis is given by the eigenvector corresponding to the eigenvalue 1, that is the vector $(0, 1, \frac{4}{3})$. The plane of rotation is the real plane in $\mathbb{R}^3$, which is given by the intersection of two dimensional complex space in $\mathbb{C}^3$ generated by the remaining eigenvectors with $\mathbb{R}^3$. It is the plane

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 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 $\mathbb{K}^n \to \mathbb{K}^n$ 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 $f : V \to V$ 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 $f : V \to V$, 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:

$\operatorname{span}\{(1,0,0),(0,-4,3)\}$ (the first generator is the (real multiple of) $v_2 + v_3$, the other one is the (real multiple of) $i(v_2 - v_3)$, see 2.4.7). We can 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).

It remains to determine the direction of the rotation. First, recall that the meaning of the direction of the rotation changes when we change the orientation of the axis (it has no meaning to speak of the direction of the rotation if we do not have an orientation of the axis). Using the ideas from the proof of the theorem 2.4.7, we see that the given matrix acts by rotating by $\arccos(\frac{3}{5})$) in the positive sense in the plane given by the basis $((1,0,0),(0,-\frac{4}{5},\frac{3}{5}))$. The first vector of the basis is the imaginary part of the eigenvector associated with the eigenvalue $\frac{3}{5} + \frac{4}{5}i$, the second is then the (common) real part of the eigenvectors associated with the complex eigenvalues. The order of the vectors in the basis is important (by changing their order the meaning of the direction changes). The axis of rotation is perpendicular to the plane. If we orient using the right-hand rule (the perpendicular direction is obtained by taking the product of the vectors in the basis) then the direction of the rotation agrees with the direction of rotation in the plane with the given basis. In our case we obtain by the vector product $(0,1,-1) \times (1,1,-1) = (0,-1,-1)$. It is thus a rotation through $\arccos(\frac{3}{5})$ in the positive sense about the vector $(0,-1,-1)$, that is, a rotation through $\arccos(\frac{3}{5})$ in the negative sense about the vector $(0,1,1)$. $\qquad\square$

2.D.19. Determine what linear mapping is given by the matrix

$$\begin{pmatrix} \frac{-1}{5} & \frac{3}{5} & \frac{-1}{5} \\ \frac{-8}{5} & \frac{9}{5} & \frac{2}{5} \\ \frac{8}{5} & \frac{-4}{5} & \frac{3}{5} \end{pmatrix}.$$

Solution. By already known method we find out that the matrix has the following eigenvalues and corresponding eigenvectors: 1, $(1,2,0)$; $\frac{3}{5} + \frac{4}{5}i$, 1, $(1,1+i,-1-i)$; $\frac{3}{5} - \frac{4}{5}i$, $(1,1-i,-1+i)$. Though all three eigenvectors have absolute value 1, they are not orthogonal to each other, thus the matrix is not orthogonal. Consequently it is not a matrix of rotation. Nevertheless, it is a linear mapping which is "close" to a rotation. It is a rotation in the plane given by two complex eigenvectors (but this plane is not orthogonal to the vector $(1,2,0)$, but it is preserved by the map). It remains to

Theorem. *Let $f : V \to V$ 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 $\lambda \in \mathbb{C}\backslash\mathbb{R}$, where f acts by the rotation by the angle equal to the argument of the complex number λ 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, 12.2.8). Furthermore, because the characteristic polynomial of the mapping has only real coefficients, the roots are either real or there are a pair of roots which are complex conjugates λ 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 λ. 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 λ and let u_λ 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 $W = \operatorname{span}\{u_\lambda\} \oplus \operatorname{span}\{\bar u_\lambda\}$, 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_λ

$$x_\lambda = \operatorname{Re} u_\lambda, \quad -y_\lambda = -\operatorname{Im} u_\lambda.$$

determine the direction of the rotation. First, we should recall that the meaning of the direction of the rotation changes when we change the orientation of the axis (it has no meaning to speak of the direction of the rotation if we do not have an orientation of the axis).

Using the same ideas as in the previous example, we see that the given matrix acts by rotating by $\arccos(\frac{3}{5}))$ in the positive sense in the plane given by the basis $((1, 1, -1), (0, 1,).$ The first vector of the basis is the imaginary part of the eigenvector associated with the eigenvalue $\frac{3}{5} + \frac{4}{5}i$, the second is then the (common) real part of the eigenvectors associated with the complex eigenvalues. The order of the vectors in the basis is important (by changing their order the meaning of the direction changes). The "axis" of rotation is not perpendicular to the plane, but we can orient the vectors lying in the whole half-plane using the right-hand rule (the perpendicular direction is obtained by taking the product of the vectors in the basis) then the direction of the rotation agrees with the direction of rotation in the plane with the given basis. In our case we obtain by the vector product $(0, 1, -1) \times (1, 1, -1) = (0, -1, -1)$. It is thus a rotation through $\arccos(\frac{3}{5})$ in the positive sense about the vector $(0, -1, -1)$, that is, a rotation through $\arccos(\frac{3}{5})$ in the negative sense about the vector $(0, 1, 1)$. □

2.D.20. Without any written computation determine the spectrum of the linear mapping $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by $(x_1, x_2, x_3) \mapsto (x_1 + x_3, x_2, x_1 + x_3)$. ○

2.D.21. Find the dimension of the eigenspaces of the eigenvalues λ_i of the matrix
$$\begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 5 & 2 & 3 & 0 \\ 0 & 4 & 0 & 3 \end{pmatrix}.$$

○

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

$$A \cdot x_\lambda = \alpha x_\lambda + \beta y_\lambda, \ A \cdot y_\lambda = -\alpha y_\lambda + \beta x_\lambda.$$

Because our mapping preserves lengths, the absolute value of the eigenvalue λ 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 λ. Note that the choice of the eigenvalue $\bar{\lambda}$ instead of λ leads to the same subspace with the same rotation, we would just have expressed it in the basis x_λ, y_λ, 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. □

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 ± 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:

2.4.8. Definition. The *spectrum of a linear mapping* f : $V \to V$, 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 ± 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.

E. Additional exercises for the whole chapter

2.E.1. Kirchhoff's Circuit Laws. We consider an application of Linear Algebra to analysis of electric circuits, using Ohm's law and Kirchhoff's voltage and current laws.

Consider an electric circuit as in the figure and write down the values of the currents there if you know the values

$$V_1 = 20, \quad V_2 = 120, \quad V_3 = 50, \quad R_1 = 10, \quad R_2 = 30, \quad R_3 = 4, \quad R_4 = 5, \quad R_5 = 10,$$

Notice that the quantities I_i denote the electric currents, while R_j are resistances, and V_k are voltages.
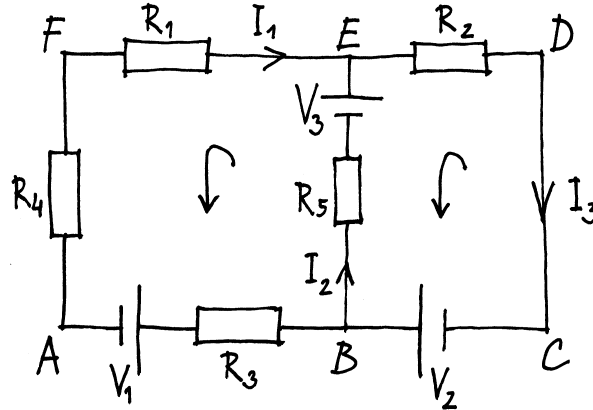
Solution. There are two closed loops, namely $ABEF$ and $EBCD$ and two branching vertices B and E of degree no less than 3. On every segment of the circuit, bounded by branching points, the electric current is constant. Set it to be I_1 on the segment $EFAB$, I_2 on EB, and I_3 on $BCDE$.

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 Kirchhhoff'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

$$-V_2 + I_3 R_2 - V_3 + R_5 I_2 = 0.$$

By combining all equations, we obtain the system

$$
\begin{array}{rcrcrcr}
I_1 & + & I_2 & - & I_3 & = & 0, \\
(R_3 + R_1 + R_4)I_1 & - & R_5 I_2 & + & & = & -V_1 - V_3, \\
& & R_5 I_2 & + & R_2 I_3 & = & V_2 + V_3.
\end{array}
$$

Substituing the prescribed values we obtain the linear system

$$
\begin{array}{rcrcrcr}
I_1 & + & I_2 & - & I_3 & = & 0, \\
19 I_1 & - & 10 I_2 & + & & = & -70, \\
& & 10 I_2 & + & 30 I_3 & = & 170.
\end{array}
$$

This has solutions $I_1 = -\frac{80}{53} \approx -1.509, \quad I_2 = \frac{219}{53} \approx 4.132, \quad I_3 = \frac{139}{53} \approx 2.623.$ □

2.E.2. The general case. In general, the method for electrical circuit analysis can be formulated along the following steps:

 i) Identify all branching vertices of the circuit, i.e vertices of degree no less than 3;

 ii) Identify all closed loops of the circuit;

 iii) Introduce variables I_k, denoting oriented currents on each segment of the circuit between two branching vertices;

iv) Write down Kirchhoff's current conservation law for each branching vertex. The total incoming current equals the total outgoing current;

v) Choose an orientation on every closed loop of the circuit and write down Kirchhoff's voltage conservation law according to the chosen orientation. If you find an electric charge of voltage V_j and you go from the short bar to the long bar, the contribution of this charge is V_j. It is $-V_j$ if you go from the long bar to the short one. If you go in the positive direction of a current I and find a resistor with resistance R_j, the contribution is $-R_j I$, and it is $R_j I$ 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.

vi) Compose the system of linear equations collecting all equations, representing Kirchhoff's current and voltage laws and solve it with respect to the variables, representing 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.
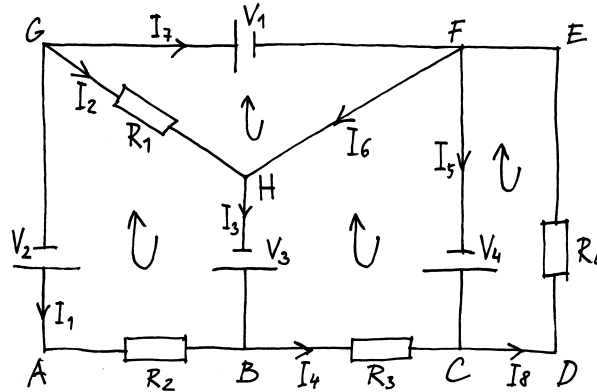
Solution.

i) The set of branching vertices is $\{B, C, F, G, H\}$.

ii) The set of closed loops is $\{ABHG, FHBC, GHF, CDEF\}$.

iii) 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 $CDEF$.

iv) Write Kirchhoff's current conservation laws for the branching vertices:

- vertex B: $I_1 + I_3 = I_4$
- vertex C: $I_4 + I_5 = I_8$
- vertex F: $I_8 = I_5 + I_6 - I_7$
- vertex G: $-I_7 = I_1 + I_2$
- vertex H: $I_2 + I_6 = I_3$

v) Write Kirchhoff's voltage conservation for each of the closed loops traversed counter-clockwise:

- loop $ABHG$: $-R_1 I_2 + V_3 + R_2 I_1 - V_2 = 0$
- loop $FHBC$: $V_4 + R_3 I_4 - V_3 = 0$
- loop GHF: $R_1 I_2 - V_1 = 0$
- loop $CDEF$: $R_4 I_8 - V_4 = 0$

Set the parameters: $R_1 = 4$, $R_2 = 7$, $R_3 = 9$, $R_4 = 12$, $V_1 = 10$, $V_2 = 20$, $, V_3 = 60$, $, V_4 = 120$, to obtain the system

$$I_1 + I_3 - I_4 = 0$$

$$I_4 + I_5 - I_8 7 = 0$$

$$I_5 + I_6 - I_7 - I_8 = 0$$

$$I_1 + I_2 + I_7 = 0$$

$$I_2 - I_3 + I_6 = 0$$
$$7I_1 - 4I_2 = -40$$
$$9I_4 = -60$$
$$4I_2 = 10$$
$$12I_8 = 120$$

with the solution set $I_1 = \frac{-30}{7}, \quad I_2 = \frac{5}{2}, \quad I_3 = \frac{-50}{21}, \quad I_4 = \frac{-20}{3}, \quad I_5 = \frac{50}{3}, \quad I_6 = \frac{-205}{42}, \quad I_7 = \frac{25}{14}, \quad I_8 = 10.$

$\square$

2.E.3. Solve the system of equations

$$
\begin{array}{rcrcrcrcrcl}
x_1 & + & x_2 & + & x_3 & + & x_4 & - & 2x_5 & = & 3, \\
 & & 2x_2 & + & 2x_3 & + & 2x_4 & - & 4x_5 & = & 5, \\
-x_1 & - & x_2 & - & x_3 & + & x_4 & + & 2x_5 & = & 0, \\
-2x_1 & + & 3x_2 & + & 3x_3 & & & - & 6x_5 & = & 2.
\end{array}
$$

Solution. The extended matrix of the system is

$$
\left(\begin{array}{ccccc|c}
1 & 1 & 1 & 1 & -2 & 3 \\
0 & 2 & 2 & 2 & -4 & 5 \\
-1 & -1 & -1 & 1 & 2 & 0 \\
-2 & 3 & 3 & 0 & -6 & 2
\end{array} \right).
$$

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

$$
\left(\begin{array}{ccccc|c}
1 & 1 & 1 & 1 & -2 & 3 \\
0 & 2 & 2 & 2 & -4 & 5 \\
0 & 0 & 0 & 2 & 0 & 3 \\
0 & 5 & 5 & 2 & -10 & 8
\end{array} \right)
\sim
\left(\begin{array}{ccccc|c}
1 & 1 & 1 & 1 & -2 & 3 \\
0 & 2 & 2 & 2 & -4 & 5 \\
0 & 0 & 0 & 2 & 0 & 3 \\
0 & 0 & 0 & -3 & 0 & -9/2
\end{array} \right).
$$

The last row is clearly a multiple of the previous, and thus we can omit it. The pivots are located in the first, second and fourth. Thus the free variables are x_3 and x_5 which we substitute by the real parameters t and s. Thus we consider the system

$$
\begin{array}{rcrcrcrcrcl}
x_1 & + & x_2 & + & t & + & x_4 & - & 2s & = & 3, \\
 & & 2x_2 & + & 2t & + & 2x_4 & - & 4s & = & 5, \\
 & & & & & & 2x_4 & & & = & 3.
\end{array}
$$

We see that $x_4 = 3/2$. The second equation gives

$$2x_2 + 2t + 3 - 4s = 5, \quad \text{that is,} \quad x_2 = 1 - t + 2s.$$

From the first we have

$$x_1 + 1 - t + 2s + t + 3/2 - 2s = 3, \quad \text{tj.} \quad x_1 = 1/2.$$

Altogether,

$$(x_1, x_2, x_3, x_4, x_5) = (1/2, 1-t+2s, t, 3/2, s), \quad t, s \in \mathbb{R}.$$

Alternatively, we can consider the extended matrix and transform it using the row transformations into the row echelon form. We arrange it so that the first non-zero number in every row is 1, and the remaining numbers in the column containing this 1 are 0. We omit the fourth equation, which is a combination of the first three. Sequentially, multiplying the second and

the third row by the number $1/2$, subtracting the third row from the second and from the first and by subtracting the second row from the first we obtain

$$\begin{pmatrix} 1 & 1 & 1 & 1 & -2 & | & 3 \\ 0 & 2 & 2 & 2 & -4 & | & 5 \\ 0 & 0 & 0 & 2 & 0 & | & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 & 1 & -2 & | & 3 \\ 0 & 1 & 1 & 1 & -2 & | & 5/2 \\ 0 & 0 & 0 & 1 & 0 & | & 3/2 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & 1 & 1 & 0 & -2 & | & 3/2 \\ 0 & 1 & 1 & 0 & -2 & | & 1 \\ 0 & 0 & 0 & 1 & 0 & | & 3/2 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & | & 1/2 \\ 0 & 1 & 1 & 0 & -2 & | & 1 \\ 0 & 0 & 0 & 1 & 0 & | & 3/2 \end{pmatrix}.$$

If we choose again $x_3 = t$, $x_5 = s$ $(t, s \in \mathbb{R})$, we obtain the general solution (2.E.3) as above. $\quad\square$

2.E.4. Find the solution of the system of linear equations given by the extended matrix

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 2 & 1 & 1 & 0 & | & 4 \\ 0 & 5 & -4 & 3 & | & 1 \\ 5 & 3 & 3 & -3 & | & 5 \end{pmatrix}.$$

Solution. We transform the given extended matrix into the row echelon form. We first copy the first three rows and into the last row we write the sum of the (2)-multiple of the first and of the (-3)-multiple of the last row. By this we obtain

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 2 & 1 & 1 & 0 & | & 4 \\ 0 & 5 & -4 & 3 & | & 1 \\ 5 & 3 & 3 & -3 & | & 5 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 5 & -4 & 3 & | & 1 \\ 0 & 6 & 1 & 14 & | & 0 \end{pmatrix}.$$

Copying the first two rows and adding a 5-multiple of the second row to the 3-multiple of the third and its 2-multiple to the fourth gives

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 5 & -4 & 3 & | & 1 \\ 0 & 6 & 1 & 14 & | & 0 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & -17 & -1 & | & 33 \\ 0 & 0 & -1 & 10 & | & 12 \end{pmatrix}.$$

Copying the first, second and fourth row, and adding the fourth to the third, yields

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & -17 & -1 & | & 33 \\ 0 & 0 & -1 & 10 & | & 12 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & -18 & 9 & | & 45 \\ 0 & 0 & -1 & 10 & | & 12 \end{pmatrix}.$$

With three more row transformations, we arrive at

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & -18 & 9 & | & 45 \\ 0 & 0 & -1 & 10 & | & 12 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & 2 & -1 & | & -5 \\ 0 & 0 & 1 & -10 & | & -12 \end{pmatrix} \sim$$

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & 1 & -10 & | & -12 \\ 0 & 0 & 2 & -1 & | & -5 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & 1 & -10 & | & -12 \\ 0 & 0 & 0 & 19 & | & 19 \end{pmatrix}.$$

The system has exactly 1 solution. We determine it by backwards elimination

$$\begin{pmatrix} 3 & 3 & 2 & 1 & | & 3 \\ 0 & -3 & -1 & -2 & | & 6 \\ 0 & 0 & 1 & -10 & | & -12 \\ 0 & 0 & 0 & 1 & | & 1 \end{pmatrix} \sim \begin{pmatrix} 3 & 3 & 2 & 0 & | & 2 \\ 0 & -3 & -1 & 0 & | & 8 \\ 0 & 0 & 1 & 0 & | & -2 \\ 0 & 0 & 0 & 1 & | & 1 \end{pmatrix} \sim$$

$$\begin{pmatrix} 3 & 3 & 0 & 0 & | & 6 \\ 0 & -3 & 0 & 0 & | & 6 \\ 0 & 0 & 1 & 0 & | & -2 \\ 0 & 0 & 0 & 1 & | & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 0 & 0 & | & 2 \\ 0 & 1 & 0 & 0 & | & -2 \\ 0 & 0 & 1 & 0 & | & -2 \\ 0 & 0 & 0 & 1 & | & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 & | & 4 \\ 0 & 1 & 0 & 0 & | & -2 \\ 0 & 0 & 1 & 0 & | & -2 \\ 0 & 0 & 0 & 1 & | & 1 \end{pmatrix}.$$

The solution is

$$x_1 = 4, \quad x_2 = -2, \quad x_3 = -2, \quad x_4 = 1.$$

$\square$

2.E.5. Find all the solutions of the homogeneous system

$$x + y = 2z + v, \quad z + 4u + v = 0, \quad -3u = 0, \quad z = -v$$

of four linear equations with 5 variables x, y, z, u, v.

Solution. We rewrite the system into a matrix such that in the first column there are coefficients of x, in the second there are coefficients of y, and so on. We put all the variables in equations to the left side. By this, we obtain the matrix

$$\begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

We add $(4/3)$-multiple of the third row to the second and subtract then the second row from the fourth to obtain

$$\begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We multiply the third row by the number $-1/3$ and add the 2-multiple of the second row to the first, which gives

$$\begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

From the last matrix, we get immediately (reading from bottom to top) $u = 0$, $z + v = 0$, $x + y + v = 0$. Letting $v = s$ and $y = t$, the complete solution is

$$(x, y, z, u, v) = (-t - s, t, -s, 0, s), \quad t, s \in \mathbb{R}.$$

which can be rewritten as

$$\begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = t \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -1 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \quad t, s \in \mathbb{R},$$

Notice that the second and the fifth column of the matrix together form a basis for the solutions. These are the columns which do not contain a leading 1 in any of its entries. $\square$

2.E.6. Determine the number of solutions for the systems

(a)

$$\begin{array}{rcrcrcr} 12x_1 & + & \sqrt{5}x_2 & + & 11x_3 & = & -9, \\ x_1 & & & - & 5x_3 & = & -9, \\ x_1 & & & + & 2x_3 & = & -7; \end{array}$$

(b)

$$\begin{array}{rcrcrcr} 4x_1 & + & 2x_2 & - & 12x_3 & = & 0, \\ 5x_1 & + & 2x_2 & - & x_3 & = & 0, \\ -2x_1 & - & x_2 & + & 6x_3 & = & 4; \end{array}$$

(c)
$$\begin{aligned}
4x_1 &+ 2x_2 &- 12x_3 &= 0, \\
5x_1 &+ 2x_2 &- x_3 &= 1, \\
-2x_1 &- x_2 &+ 6x_3 &= 0.
\end{aligned}$$

Solution. The vectors $(1, 0, -5)$, $(1, 0, 2)$ are clearly linearly independent, (they are not multiples of each other) and the vector $(12, \sqrt{5}, 11)$ cannot be their linear combination (its second coordinate is non-zero). Therefore the matrix whose rows are these three linearly independent vectors (from the left side) is invertible. Thus the system for case (a) has exactly one solution.

For cases (b) and (c), it is enough to note that

$$(4, 2, -12) = -2(-2, -1, 6).$$

In case (b) adding the first equation to the third multiplied by two gives $0 = 8$, hence there is no solution for the system. In case (c) the third equation is a multiple of the first, so the system has infinitely many distinct solutions. □

2.E.7. Find a linear system, whose set of solutions is exactly

$$\{(t + 1,\ 2t,\ 3t,\ 4t);\ t \in \mathbb{R}\}.$$

Solution. Such a system is for instance

$$2x_1 - x_2 = 2, \quad 2x_2 - x_4 = 0, \quad 4x_3 - 3x_4 = 0.$$

These solutions are satisfied for every $t \in \mathbb{R}$. The vectors

$$(2, -1, 0, 0), \quad (0, 2, 0, -1), \quad (0, 0, 4, -3)$$

giving the left-hand sides of the equations are linearly independent (the set of solutions contains a single parameter). □

2.E.8. Solve the system of homogeneous linear equations given by the matrix

$$\begin{pmatrix}
0 & \sqrt{2} & \sqrt{3} & \sqrt{6} & 0 \\
2 & 2 & \sqrt{3} & -2 & -\sqrt{5} \\
0 & 2 & \sqrt{5} & 2\sqrt{3} & -\sqrt{3} \\
3 & 3 & \sqrt{3} & -3 & 0
\end{pmatrix}.$$

○

2.E.9. Determine all solutions of the system

$$\begin{aligned}
x_2 & & &+ x_4 &= 1, \\
3x_1 &- 2x_2 &- 3x_3 &+ 4x_4 &= -2, \\
x_1 &+ x_2 &- x_3 &+ x_4 &= 2, \\
x_1 & &- x_3 & &= 1.
\end{aligned}$$

○

2.E.10. Solve

$$\begin{aligned}
3x &- 5y &+ 2u &+ 4z &= 2, \\
5x &+ 7y &- 4u &- 6z &= 3, \\
7x &- 4y &+ &+ 3z &= 4, \\
x &+ 6y &- 2u &- 5z &= 2
\end{aligned}$$

○

2.E.11. Determine whether or not the system of linear equations

$$\begin{aligned}
3x_1 &+ 3x_2 &+ x_3 &= 1, \\
2x_1 &+ 3x_2 &- x_3 &= 8, \\
2x_1 &- 3x_2 &+ x_3 &= 4, \\
3x_1 &- 2x_2 &+ x_3 &= 6
\end{aligned}$$

of three variables x_1, x_2, x_3 has a solution.

○

2.E.12. Determine the number of solutions of the system of 5 linear equations

$$A^T \cdot x = (1, 2, 3, 4, 5)^T,$$

where

$$x = (x_1, x_2, x_3)^T \quad \text{and} \quad A = \begin{pmatrix} 3 & 1 & 7 & 5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 2 & 1 & 4 & 3 & 0 \end{pmatrix}.$$

Repeat the question for the system

$$A^T \cdot x = (1, 1, 1, 1, 1)^T$$

$\bigcirc$

2.E.13. Depending on the parameter $a \in \mathbb{R}$, determine the solution of the system of linear equations

$$\begin{array}{rcrcrcl} ax_1 & + & 4x_2 & +2 & x_3 & = & 0, \\ 2x_1 & + & 3x_2 & - & x_3 & = & 0. \end{array}$$

$\bigcirc$

2.E.14. Depending on the parameter $a \in \mathbb{R}$, determine the number of solutions of the system

$$\begin{pmatrix} 4 & 1 & 4 & a \\ 2 & 3 & 6 & 8 \\ 3 & 2 & 5 & 4 \\ 6 & -1 & 2 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 3 \\ -3 \end{pmatrix}.$$

$\bigcirc$

2.E.15. Decide whether or not there is a system of homogeneous linear equations of three variables whose set of solutions is exactly

(a) $\{(0, 0, 0)\}$;

(b) $\{(0, 1, 0), (0, 0, 0), (1, 1, 0)\}$;

(c) $\{(x, 1, 0); \ x \in \mathbb{R}\}$;

(d) $\{(x, y, 2y); \ x, y \in \mathbb{R}\}$.

$\bigcirc$

2.E.16. Solve the system of linear equations, depending on the real parameters a, b.

$$\begin{array}{rcl} x + 2y + bz & = & a \\ x - y + 2z & = & 1 \\ 3x - y & = & 1. \end{array}$$

$\bigcirc$

2.E.17. Using the inverse matrix, compute the solution of the system

$$\begin{array}{rcl} x_1 + x_2 + x_3 + x_4 & = & 2, \\ x_1 + x_2 - x_3 - x_4 & = & 3, \\ x_1 - x_2 + x_3 - x_4 & = & 3, \\ x_1 - x_2 - x_3 + x_4 & = & 5. \end{array}$$

$\bigcirc$

2.E.18. For what values of parameters $a, b \in \mathbb{R}$ has the system of linear equations

$$
\begin{array}{rcrcrcl}
x_1 & - & ax_2 & - & 2x_3 & = & b, \\
x_1 & + & (1-a)x_2 & & & = & b - 3, \\
x_1 & + & (1-a)x_2 & + & ax_3 & = & 2b - 1
\end{array}
$$

(a) exactly one solution;

(b) no solution;

(c) at least 2 solutions? (i.e. infinitely many solutions)

Solution. We rewrite it, as usual, in the extended matrix, and transform:

$$
\begin{pmatrix}
1 & -a & -2 & b \\
1 & 1-a & 0 & b-3 \\
1 & 1-a & a & 2b-1
\end{pmatrix}
\sim
\begin{pmatrix}
1 & -a & -2 & b \\
0 & 1 & 2 & -3 \\
0 & 1 & a+2 & b-1
\end{pmatrix}
$$

$$
\sim
\begin{pmatrix}
1 & -a & -2 & b \\
0 & 1 & 2 & -3 \\
0 & 0 & a & b+2
\end{pmatrix}.
$$

At the first step we subtract the first row from the second and the third; and at the second step we subtract the second 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 $a = 0$, $b = -2$ the solutions are

$$
(x_1,\, x_2,\, x_3) = (-2 + 2t,\, -3 - 2t,\, t)\,, \quad t \in \mathbb{R}
$$

and for $a \neq 0$ the unique solution is the triple

$$
\left(\frac{-3a^2 - ab - 4a + 2b + 4}{a},\, -\frac{2b + 3a + 4}{a},\, \frac{b+2}{a} \right).
$$

$\square$

2.E.19. Let

$$
A = \begin{pmatrix} 4 & 5 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad
x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad
b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.
$$

Find real numbers b_1, b_2, b_3 such that the system of linear equations $A \cdot x = b$ has:

(a) infinitely many solutions;

(b) unique solution;

(c) no solution;

(d) exactly four solutions.

Solution. Since the first row is the sum of the other two, it is enough to choose $b_1 = b_2 + b_3$ in case a) and $b_1 \neq b_2 + b_3$ in case c). Variant b) cannot occur, since the matrix A is not invertible. As long as we work over reals, there cannot be any finite number of solutions, except zero or one. Thus d) is impossible. $\square$

2.E.20. Factor the following permutations into a product of transpositions:

i) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$,

ii) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 4 & 1 & 2 & 5 & 8 & 3 & 7 \end{pmatrix}$,

iii) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 6 & 1 & 10 & 2 & 5 & 9 & 8 & 3 & 7 \end{pmatrix}$.

2.E.21. Determine the parity of the given permutations:

i) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 5 & 6 & 4 & 1 & 2 & 3 \end{pmatrix}$,

ii) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 7 & 1 & 2 & 3 & 8 & 4 & 5 \end{pmatrix}$,

iii) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 7 & 1 & 10 & 2 & 5 & 4 & 9 & 3 & 6 \end{pmatrix}$.

2.E.22. Find the algebraically adjoint matrix F^* for

$$F = \begin{pmatrix} \alpha & \beta & 0 \\ \gamma & \delta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \alpha, \beta, \gamma, \delta \in \mathbb{R}.$$

2.E.23. Calculate the algebraically adjoint matrix for the matrices

(a) $\begin{pmatrix} 3 & -2 & 0 & -1 \\ 0 & 2 & 2 & 1 \\ 1 & -2 & -3 & -2 \\ 0 & 1 & 2 & 1 \end{pmatrix}$, (b) $\begin{pmatrix} 1+i & 2i \\ 3-2i & 6 \end{pmatrix}$,

where i denotes the imaginary unit.

2.E.24. Is the set $V = \{(1, x); \, x \in \mathbb{R}\}$ with operations

$$\oplus : V \times V \to V, \quad (1, y) \oplus (1, z) = (1, z + y) \quad \text{for all } z, y \in \mathbb{R}$$

$$\odot : \mathbb{R} \times V \to V, \quad z \odot (1, y) = (1, y \cdot z) \quad \text{for all } z, y \in \mathbb{R}$$

a vector space?

2.E.25. Express the vector $(5, 1, 11)$ as a linear combination of the vectors $(3, 2, 2), (2, 3, 1), (1, 1, 3)$, that is, find numbers $p, q, r \in \mathbb{R}$, for which

$$(5, 1, 11) = p\,(3, 2, 2) + q\,(2, 3, 1) + r\,(1, 1, 3).$$

2.E.26. In $\mathbb{R}^3$, determine the matrix of rotation through the angle $120°$ in the positive sense about the vector $(1, 0, 1)$

2.E.27. In the vector space $\mathbb{R}^3$, determine the matrix of the orthogonal projection onto the plane $x + y - 2z = 0$.

2.E.28. In the vector space $\mathbb{R}^3$, determine the matrix of the orthogonal projection on the plane $2x - y + 2z = 0$.

2.E.29. Determine whether the subspaces $U = \langle (2, 1, 2, 2) \rangle$ and $V = \langle (-1, 0, -1, 2), (-1, 0, 1, 0), (0, 0, 1, -1) \rangle$ of the space $\mathbb{R}^4$ are orthogonal. If they are, is $\mathbb{R}^4 = U \oplus V$, that is, is $U^\perp = V$?

2.E.30. Let p be a given line:

$$p : [1, 1] + (4, 1)t, \, t \in \mathbb{R}$$

Determine the parametric expression of all lines q that pass through the origin and have deflection $60°$ with the line p.

2.E.31. Depending on the parameter $t \in \mathbb{R}$, determine the dimension of the subspace U of the vector space $\mathbb{R}^3$, if U is generated by the vectors

(a) $u_1 = (1, 1, 1)$, $\quad u_2 = (1, t, 1)$, $\quad u_3 = (2, 2, t)$;
(b) $u_1 = (t, t, t)$, $\quad u_2 = (-4t, -4t, 4t)$, $\quad u_3 = (-2, -2, -2)$.

○

2.E.32. Construct an orthogonal basis of the subspace

$$\langle\, (1, 1, 1, 1),\ (1, 1, 1, -1),\ (-1, 1, 1, 1) \,\rangle$$

of the space $\mathbb{R}^4$. ○

2.E.33. In the space $\mathbb{R}^4$, find an orthogonal basis of the subspace of all linear combinations of the vectors $(1, 0, 1, 0)$, $(0, 1, 0, -7)$, $\quad (4, -2, 4, 14)$.

   Find an orthogonal basis of the subspace generated by the vectors $(1, 2, 2, -1)$, $\quad (1, 1, -5, 3)$, $\quad (3, 2, 8, -7)$. ○

2.E.34. For what values of the parameters $a, b \in \mathbb{R}$ are the vectors

$$(1, 1, 2, 0, 0), \quad (1, -1, 0, 1, a), \quad (1, b, 2, 3, -2)$$

in the space $\mathbb{R}^5$ pairwise orthogonal? ○

2.E.35. In the space $\mathbb{R}^5$, consider the subspace generated by the vectors $(1, 1, -1, -1, 0), (1, -1, -1, 0, -1), (1, 1, 0, 1, 1), (-1, 0, -1, 1, 1)$. Find a basis for its orthogonal complement. ○

2.E.36. Describe the orthogonal complement of the subspace V of the space $\mathbb{R}^4$, if V is generated by the vectors $(-1, 2, 0, 1)$, $(3, 1, -2, 4), (-4, 1, 2, -4), (2, 3, -2, 5)$. ○

2.E.37. In the 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 system of equations $x_1 - x_3 = 0$, $x_1 - x_2 + x_3 - x_4 + x_5 = 0$.

○

2.E.38. In the space $\mathbb{R}^4$, let

$$(1, -2, 2, 1), \quad (1, 3, 2, 1)$$

be given vectors. Extend these two vectors into an orthogonal basis of the whole $\mathbb{R}^4$. (You can do this in any way you wish, for instance by using the Gram-Schmidt orthogonalization process.) ○

2.E.39. Define an inner product on the vector space of the matrices from the previous exercise. Compute the norm of the matrix from the previous exercise, induced by the product you have defined. ○

2.E.40. Find a basis for the vector space of all antisymmetric real square matrices of the type 4×4. Consider the standard inner product in this basis and using this inner product, express the size of the matrix

$$\begin{pmatrix} 0 & 3 & 1 & 0 \\ -3 & 0 & 1 & 2 \\ -1 & -1 & 0 & 2 \\ 0 & -2 & -2 & 0 \end{pmatrix}$$

.
○

2.E.41. Find the eigenvalues and the associated eigenspaces of eigenvectors of the matrix:

$$A = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ 2 & -2 & 2 \end{pmatrix}.$$

Solution. The characteristic polynomial of the matrix is $\lambda^3 - 6\lambda^2 + 12\lambda - 8$, which is $(\lambda - 2)^3$. The number 2 is thus an eigenvalue with algebraic multiplicity three. Its geometric multiplicity is either one, two or three. We determine the vectors associated to this eigenvalue as the solutions of the system

$$(A - 2E)\mathbf{x} = \begin{matrix} -x_1 & +x_2 & = & 0, \\ -x_1 & +x_2 & = & 0, \\ 2x_1 & -2x_2 & = & 0. \end{matrix}$$

Its solutions form the two-dimensional space $\langle (1, -1, 0), (0, 0, 1) \rangle$. Thus the eigenvalue 2 has algebraic multiplicity 3 and geometric multiplicity 2.

□

2.E.42. Determine the eigenvalues of the matrix

$$\begin{pmatrix} -13 & 5 & 4 & 2 \\ 0 & -1 & 0 & 0 \\ -30 & 12 & 9 & 5 \\ -12 & 6 & 4 & 1 \end{pmatrix}.$$

○

2.E.43. Given that the numbers $1, -1$ are eigenvalues of the matrix

$$A = \begin{pmatrix} -11 & 5 & 4 & 1 \\ -3 & 0 & 1 & 0 \\ -21 & 11 & 8 & 2 \\ -9 & 5 & 3 & 1 \end{pmatrix},$$

find all solutions of the characteristic equation $|A - \lambda E| = 0$. Hint: if you denote all the roots of the polynomial $|A - \lambda E|$ by $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, then

$$|A| = \lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \lambda_4, \quad and \quad \operatorname{tr} A = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4.$$

○

2.E.44. Find a four-dimensional matrix with eigenvalues $\lambda_1 = 6$ and $\lambda_2 = 7$ such that the multiplicity of λ_2 as a root of the characteristic polynomial is three, and that

(a) the dimension of the subspace of eigenvectors of λ_2 is 3;

(b) the dimension of the subspace of eigenvectors of λ_2 is 2;

(c) the dimension of the subspace of eigenvectors of λ_2 is 1;

○

2.E.45. Find the eigenvalues and the eigenvectors of the matrix:

$$\begin{pmatrix} -1 & -\frac{5}{6} & \frac{5}{3} \\ 0 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{1}{6} & -\frac{4}{3} \end{pmatrix}.$$

2.E.46. Determine the characteristic polynomial $|A - \lambda E|$, eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 4 & -1 & 6 \\ 2 & 1 & 6 \\ 2 & -1 & 8 \end{pmatrix}.$$

respectively.

Solutions to the exercises

2.A.11. There is only one such matrix X, and it is

$$\begin{pmatrix} 18 & -32 \\ 5 & -8 \end{pmatrix}.$$

2.A.13. $A^{-1} = \begin{pmatrix} 1 & 10 & -4 \\ 1 & 12 & -5 \\ 0 & 5 & -2 \end{pmatrix}.$

2.A.14.

$$A^5 = \begin{pmatrix} 122 & -121 & 121 \\ -121 & 122 & -121 \\ 0 & 0 & 1 \end{pmatrix}, \quad A^{-3} = \tfrac{1}{27} \begin{pmatrix} 14 & 13 & -13 \\ 13 & 14 & 13 \\ 0 & 0 & 27 \end{pmatrix}.$$

2.A.15. $\begin{pmatrix} 2 & -3 & 0 & 0 & 0 \\ -5 & 8 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -5 & 2 \\ 0 & 0 & 0 & 3 & -1 \end{pmatrix}.$

2.A.16. $C^{-1} = \tfrac{1}{2} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$

2.A.17. In the first case we have

$$A^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 3 & -i \\ i & 1 \end{pmatrix};$$

in the second

$$A^{-1} = \begin{pmatrix} 14 & 8 & 5 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

2.C.9. $(2 + \tfrac{1}{\sqrt{3}}, 2 - \tfrac{1}{\sqrt{3}}).$

2.C.10. The vectors are dependent whenever at least one of the conditions

$$a = b = 1, \qquad a = c = 1, \qquad b = c = 1$$

is satisfied.

2.C.11. Vectors are linearly independent.

2.C.12. It suffices to add for instance the polynomial x.

2.D.5. $\cos = \tfrac{\sqrt{2}}{\sqrt{3}}.$

2.D.12. Je $| A - \lambda E | = -\lambda^3 + 12\lambda^2 - 47\lambda + 60$, . $\lambda_1 = 3$, $\lambda_2 = 4$, $\lambda_3 = 5$.

2.D.20. The solution is the sequence $0, 1, 2$.

2.D.21. The dimension is 1 for $\lambda_1 = 4$ and 2 for $\lambda_2 = 3$.

2.E.8. The solutions are all scalar multiples of the vector

$$\left(1 + \sqrt{3}, \ -\sqrt{3}, \ 0, \ 1, \ 0 \right).$$

2.E.9. $x_1 = 1 + t$, $\quad x_2 = \tfrac{3}{2}$, $\quad x_3 = t$, $\quad x_4 = -\tfrac{1}{2}$, $\qquad t \in \mathbb{R}$.

2.E.10. The system has no solution.

2.E.11. The system has a solution, because

$$3 \cdot \begin{pmatrix} 3 \\ 2 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3 \\ -3 \\ -2 \end{pmatrix} - 5 \cdot \begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \\ 4 \\ 6 \end{pmatrix}.$$

2.E.12. The system of linear equations

$$
\begin{array}{rcrcl}
3x_1 & + & 2x_3 & = & 1,\\
x_1 & + & x_3 & = & 2,\\
7x_1 & + & 4x_3 & = & 3,\\
5x_1 & + & 3x_3 & = & 4,\\
x_2 & & & = & 5
\end{array}
$$

has no solution, while the system

$$
\begin{array}{rcrcl}
3x_1 & + & 2x_3 & = & 1,\\
x_1 & + & x_3 & = & 1,\\
7x_1 & + & 4x_3 & = & 1,\\
5x_1 & + & 3x_3 & = & 1,\\
x_2 & & & = & 1
\end{array}
$$

has a unique solution $x_1 = -1$, $x_2 = 1$, $x_3 = 2$.

2.E.13. The set of all solutions is given by

$$\{(-10t,\ (a+4)t,\ (3a-8)t)\,;\ t \in \mathbb{R}\}.$$

2.E.14. For $a = 0$, the system has no solution. For $a \neq 0$ the system has infinitely many solutions.

2.E.15. The correct answers are „yes", „no", „no" and „yes" respectively.

2.E.16. i) If $b \neq -7$, then $x = z = (2+a)/(b+7)$, $y = (3a-b-1)/(b+7)$. ii) If $b = -7$ and $a \neq -2$, then there is no solution. iii) If $a = -2$ and $b = -7$ then the solution is $x = z = t$, $y = 3t - 1$, for any t.

2.E.17.

$$
\begin{pmatrix}
1 & 1 & 1 & 1\\
1 & 1 & -1 & -1\\
1 & -1 & 1 & -1\\
1 & -1 & -1 & 1
\end{pmatrix}^{-1}
\frac{1}{4}
\begin{pmatrix}
1 & 1 & 1 & 1\\
1 & 1 & -1 & -1\\
1 & -1 & 1 & -1\\
1 & -1 & -1 & 1
\end{pmatrix}.
$$

We can then easily obtain

$$x_1 = \frac{13}{4}, \quad x_2 = -\frac{3}{4}, \quad x_3 = -\frac{3}{4}, \quad x_4 = \frac{1}{4}.$$

2.E.20. i) $(1,7)(2,6)(5,3)$, ii) $(1,6)(6,8)(8,7)(7,3)(2,4)$, iii) $(1,4)(4,10)(10,7)(7,9)(9,3)(2,6)(6,5)$

2.E.21. i) 17 inversions, odd, ii) 12 inversions, even iii) 25 inversions, odd

2.E.22. From the knowledge of the inverse matrix F^{-1} we obtain

$$
F^* = (\alpha\delta - \beta\gamma)\, F^{-1} =
\begin{pmatrix}
\delta & -\beta & 0\\
-\gamma & \alpha & 0\\
0 & 0 & \alpha\delta - \beta\gamma
\end{pmatrix},
$$

for any $\alpha, \beta, \gamma, \delta \in \mathbb{R}$.

2.E.23. The matrices are

(a) $\begin{pmatrix} 1 & 1 & -2 & -4\\ 0 & 1 & 0 & -1\\ -1 & -1 & 3 & 6\\ 2 & 1 & -6 & -10 \end{pmatrix}$, (b) $\begin{pmatrix} 6 & -2i\\ -3+2i & 1+i \end{pmatrix}$.

2.E.24. It is easy to check that it is a vector space. The first coordinate does not affect the results of the operations – it is just the vector space $(\mathbb{R}, +, \cdot)$ written in a different way.

2.E.25. There is a unique solution

$$p = 2, \qquad q = -2, \qquad r = 3.$$

2.E.26.

$$
\begin{pmatrix}
1/4 & -\sqrt{6}/4 & 3/4\\
\sqrt{6}/4 & -1/2 & -\sqrt{6}/4\\
3/4 & \sqrt{6}/4 & 1/4
\end{pmatrix}
$$

2.E.27.

$$
\begin{pmatrix}
5/6 & -1/6 & 1/3\\
-1/6 & 5/6 & 1/3\\
1/3 & 1/3 & 1/3
\end{pmatrix}.
$$

2.E.28.
$$\begin{pmatrix} 5/9 & 2/9 & -4/9 \\ 2/9 & 8/9 & 2/9 \\ -4/9 & 2/9 & 5/9 \end{pmatrix}$$

2.E.29. The vector that determines the subspace U is perpendicular to each of the three vectors that generate V. The subspaces are thus orthogonal. But it is not true that $\mathbb{R}^4 = U \oplus V$. The subspace V is only two-dimensional, because
$$(-1, 0, -1, 2) = (-1, 0, 1, 0) - 2(0, 0, 1, -1).$$

2.E.30.
$$q_1 : (2 - \frac{\sqrt{3}}{2}, 2\sqrt{3} + \frac{1}{2})t, \quad q_2 : (2 + \frac{\sqrt{3}}{2}, -2\sqrt{3} + \frac{1}{2})t.$$

2.E.31. In the first case we have $\dim U = 2$ for $t \in \{1, 2\}$, otherwise we have $\dim U = 3$. In the second case we have $\dim U = 2$ for $t \neq 0$ and $\dim U = 1$ for $t = 0$.

2.E.32. Using the Gram-Schmidt orthogonalization process we can obtain the result
$$((1, 1, 1, 1), (1, 1, 1, -3), (-2, 1, 1, 0)).$$

2.E.33. We have for instance the orthogonal bases
$$((1, 0, 1, 0), (0, 1, 0, -7))$$
for the first part, and
$$((1, 2, 2, -1), (2, 3, -3, 2), (2, -1, -1, -2)).$$
for the second part.

2.E.34. The solution is $a = 9/2$, $b = -5$, because
$$1 + b + 4 + 0 + 0 = 0, \quad 1 - b + 0 + 3 - 2a = 0.$$

2.E.35. The basis must contain a single vector. It is
$$(3, -7, 1, -5, 9).$$
(or any non-zero scalar multiple thereof)

2.E.36. The orthogonal complement $V^\perp$ is the set of all scalar multiples of the vector $(4, 2, 7, 0)$.

2.E.37.
(a) $W^\perp = \langle (1, 0, -1, 1, 0), (1, 3, 2, 1, -3) \rangle$;
(b) $W^\perp = \langle (1, 0, -1, 0, 0), (1, -1, 1, -1, 1) \rangle$.

2.E.38. There are infinitely many possible extensions, of course. A very simple one is
$$(1, -2, 2, 1), \quad (1, 3, 2, 1), \quad (1, 0, 0, -1), \quad (1, 0, -1, 1).$$

2.E.39. For instance, one can use the inner product that follows from the isomorphism of the space of all real 3×3 matrices with the space $\mathbb{R}^9$. If we use the product from $\mathbb{R}^9$, we obtain an inner product that assigns to two matrices the sum of products of two corresponding elements. For the given matrix we obtain
$$\left\| \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix} \right\| = \sqrt{ \left\langle \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix} \right\rangle } = \sqrt{1^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 1^2 + (-2)^2 + (-3)^2} = \sqrt{23}.$$

2.E.40.

2.E.42. The matrix has only one eigenvalue, namely -1, since the characteristic polynomial is $(\lambda + 1)^4$.

2.E.43. The root -1 of the polynomial $|A - \lambda E|$ has multiplicity three.

2.E.44. Possible examples are,

$$\text{(a)} \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix}; \quad \text{(b)} \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 1 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix};$$

$$\text{(c)} \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 1 & 0 \\ 0 & 0 & 7 & 1 \\ 0 & 0 & 0 & 7 \end{pmatrix}.$$

2.E.45. There is a triple eigenvalue -1. The corresponding eigenspace is $\langle(1,0,0),(0,2,1)\rangle$.

2.E.46. The characteristic polynomial is $-(\lambda-2)^2(\lambda-9)$, that is, the eigenvalues are 2 and 9 with associated eigenvectors
$$(1,2,0),(-3,0,1) \quad a \quad (1,1,1)$$

Linear models and matrix calculus

where are the matrices useful?
– basically almost everywhere...

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 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 Jodan canonical form, decompositions, and pseudo-inverses of matrices). The reader should feel free to move forward if getting lost.

A. Linear optimization

Let us start with an example of a very simple problem:

3.A.1. A company manufactures bolts and nuts. Nuts and bolts are moulded – moulding a box of bolts takes one minute, 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. How many boxes of nuts and bolts should be manufactured in order to have maximal profit?

Solution. Write the given data into a table:

	Bolts 1 box	Nuts 1 box	Capacity
Mould	1 min./box	2 min./box	2 hours
Box	1 min./box	4 min./box	3 hours
Profit	$4/box	$6/box	

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 φ and coordinate expression of the vector b. We have mastered such problems in the previous chapter. Now we

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:

$$x_1 + 2x_2 \leq 120$$
$$x_1 + 4x_2 \leq 180$$
$$x_1 \geq x_2 + 90$$
$$x_1 \leq 110$$

The objective function (the function that gives the profit for given number of manufactured nuts and bolts) is $4x_1 + 6x_2$. The previous system of inequalities defines a region in $\mathbb{R}^2$. Optimisation of the profit means finding in this region the point (points) in which the objective function has the maximum value, that is, to find the largest k such that the line $4x_1 + 6x_2 = k$ has a non-empty intersection with the given region. Graphically, we can find the solution for example by placing the line p into the plane such that it satisfies the equation $4x_1 + 6x_2 = 0$ and start moving it "upwards" as long as it has some intersection with the area. It is clear that the last intersection is either a point or a line parallel to p forming a border of the region. Thus we obtain (see the figure) the point $x_1 = 110$ and $x_2 = 5$. Maximum possible income is thus $4 \cdot 110 + 6 \cdot 5 = \$470$.
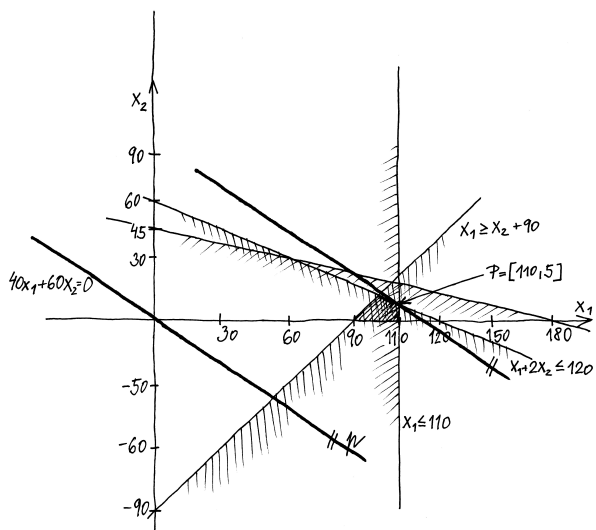
***3.A.2.* Minimisation of costs for feeding.** A stable in Nišovice u Volyně buys fodder for winter: hay and oats. The

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 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

$$(1) \qquad x_1 + x_2 \leq L$$
$$(2) \qquad w_1 x_1 + w_2 x_2 \leq W$$
$$(3) \qquad b_1 x_1 + b_2 x_2 \leq B.$$

The total net income of the painter, which is the following linear form h,

$$h(x_1, x_2) = c_1 x_1 + c_2 x_2,$$

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.
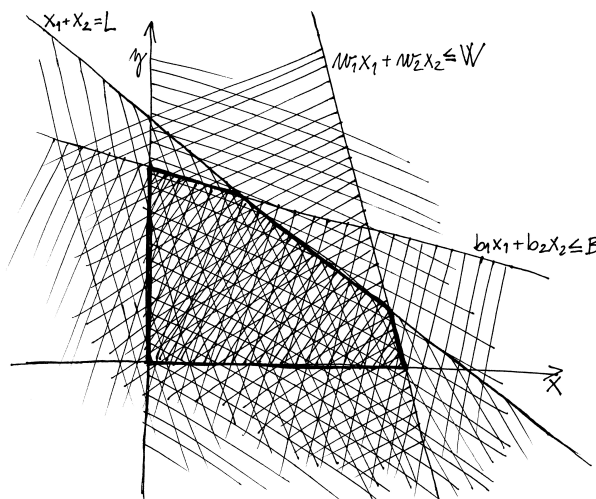
the axis in the diagram should be called x_1 and x_2! Add the line of constant value for h, best through one of the vertices with hand-written description "optimal constant value of h

nutritional values of the fodder and required daily portions for one foal are given in the table:

g/kg	Hay	Oats	Requirements
Dry basis	841	860	≥ 6300 g
Digestible nitrogen stuff	53	123	≥ 1150 g
Starch	0.348	0.868	≤ 5.35 g
Calcium	6	1.6	≥ 30 g
Phosphate	2.8	3.5	≤ 44 g
Natrium	0.2	1.4	$\simeq 7$ g
Cost	1.80	1.60	

Every foal must obtain in its daily meal at least $2\,\mathrm{kg}$ of oats. The average cost (counting the payment for the transportation) is €1.80 per $1\,\mathrm{kg}$ of hay and €1.60 per $1\,\mathrm{kg}$ of oats. Compose a daily diet for one foal which has minimum costs. ◯

The previous three examples could be solved by drawing the diagram and checking all the vertices on the boundary of the polygonal area $M \subset \mathbb{R}^2$. Morevoer, we know that the maximum will be at one of the extremes in the direction of the normal to the defining line from the linear cost function h.

But the principle works in higher dimensions as well. If there is a function $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(x_1, \ldots, x_n) = c_0 + c_1 x_1 + \cdots + c_n x_n$ (we call it the objective function), its values in the points $A = (a_1, \ldots, a_n)$ and $B = A + u = (a_1 + u_1, \ldots, a_n + u_n)$ differ by $f(A - B) = f(u) = f(u_1, \ldots, u_n) = c_1 u_1 + \cdots + c_n u_n$, which is the scalar product of the vectors $(c_1, ..., c_n)$ and $(u_1, ..., u_n)$. The relation between the scalar product and the cosine of the angle between vectors ensures that the given function f defines a hypersurface in $\mathbb{R}^n$ with normal $(c_1, \ldots, c_n)$. This hypersurface 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 in 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).

This observation leads to an algorithm for finding the extremal values of the linear objective function f on the set M of admissible points defined by linear inequalities.

We shall deal with the standard problem of linear programming. That is, we want to maximize the linear function $h = c_1 x_1 + \cdots + c_n x_n$ on the set M given by $Ax \leq b$ and $x \geq 0$ (here the inequality between vectors means the inequality between all their individual components). As explained in 3.1.6 we may add slack variables x_s, one for each equation.

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 $h : V \to \mathbb{R}$ (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

$$t \mapsto h(u + t\,v) = h(u) + t\,h(v).$$

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*.[1] 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.

---

[1]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.

Thus we may restrict ourselves to the problem of maximizing h on a vector space of solutions of systems of linear equations with the additional condition that all the values of the coordinates must be non-negative.

If there are more general inequalities, we can always change them into our form by multiplying them with -1 and minimization of value of h corresponds to maximization of $-h$.

As explained in more details in 3.1.1 and 3.1.6, we add the first row of coefficients of h (with minus signs) and use the simplex tableau:

$$\begin{array}{ccc|c} -c_1 & \ldots & -c_n & 0 \\ \hline a_{11} & \ldots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \ldots & a_{mn} & b_n \end{array}$$

We start the algorithm if we find m columns (here m is the number of equations in the problem) such that Gauss elimination for these columns leads to a unit submatrix in A and positive values at the positions of all b_i. The coordinates corresponding to these columns and values 1 in them are called the basic coordinates. We restrict ourselves to the cases where all b_i are nonnegative in the original problem and then we choose all the slack variables as the basic ones and the initialization of the algorithm is done.

Next we move in the following iterated steps (compare the more theoretical explanation in 3.1.6):

We choose the first column from the left having a non-positive value in the first row. In this column (let it be the j-th column), we pick up the positive entry a_{ij} in A which provides the minimal relation b_i/a_{ij} (we call this entry the *pivot*). Finally we eliminate the entire chosen column with the help of the chosen a_i. This means we achieve by elementary row transformations that the j-th column contains the value 1 at it's i-th row, with all other values vanishing.

We explain the procedure by a example:

3.A.3. Minimize the function $-3x - y - 2z$ under the conditions $x, y, z \geq 0$ and

$$\begin{array}{rcrcrcr} x & - & y & + & z & \geq & -4, \\ 2x & & & + & z & \leq & 3, \\ x & + & y & + & 3z & \leq & 8. \end{array}$$

Solution. First we multiply the objective function and the first inequality by -1. We get the equivalent task of maximizing

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.

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 a 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. In paragraph 4.1.11 on page 239 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.

the function $3x + y + 2z$ under the conditions

$$
\begin{array}{rcrcrcl}
-x & + & y & - & z & \leq & 4, \\
2x & & & + & z & \leq & 3, \\
x & + & y & + & 3z & \leq & 8.
\end{array}
$$

Introducing the non-negative slack variables u, v, w, we obtain the tableau with the objective function $3x + y + 2z + 0 \cdot u + 0 \cdot v + 0 \cdot w$:

-3	-1	-2	0	0	0	0
-1	1	-1	$\boxed{1}$	0	0	4
$\circled{2}$	0	1	0	$\boxed{1}$	0	3
1	1	3	0	0	$\boxed{1}$	8

Since the right-hand column is non-negative, setting $u = 4$, $v = 3$, $w = 8$, $x = y = z = 0$ provides an admissible solution to the system which corresponds to the choice of the basic variables u, v, w, and the algorithm may begin.

The first column is already with a negative entry in the first row, so we choose this one. We have circled the pivot, i.e. the value two there (we compare the relations of the elements and those in the last column, i.e. $\frac{3}{2}$ and $\frac{8}{1}$, and take the minimal one, since we need to keep the last column positive during the elimination). Next we eliminate the first column with the help of the pivot (we multiply the third row by $\frac{1}{2}$, and subtract its reasonable multiples from the other rows, not forgetting the first row of the tableau, so that only zero entries remain there):

0	-1	$-\frac{1}{2}$	0	$\frac{3}{2}$	0	$\frac{9}{2}$
0	$\circled{1}$	$-\frac{1}{2}$	$\boxed{1}$	$\frac{1}{2}$	0	$\frac{11}{2}$
$\boxed{1}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{3}{2}$
0	1	$\frac{5}{2}$	0	$-\frac{1}{2}$	$\boxed{1}$	$\frac{13}{2}$

Now the basic variables are $x = 3/2$, $u = 11/2$, $w = 13/2$, which reflects the fact that we moved as much from the former slack variable v to the new basic variable x as possible. This increased the value of the objective function, which we may read in the right top corner of the tableau.

Next, we choose the pivot from the second column and the above rule yields the first row in A ($\frac{11}{2} < \frac{13}{2}$). We have already circled the 1 in the tableau above. We eliminate:

0	0	-1	1	2	0	10
0	$\boxed{1}$	$-\frac{1}{2}$	1	$\frac{1}{2}$	0	$\frac{11}{2}$
$\boxed{1}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{3}{2}$
0	0	$\circled{3}$	-1	-1	$\boxed{1}$	1

Specifically, in our problem of the black&white painter from 3.1.1, the system of linear equations looks like this:

$$
\begin{pmatrix}
1 & -c_1 & -c_2 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & w_1 & w_2 & 0 & 1 & 0 \\
0 & b_1 & b_2 & 0 & 0 & 1
\end{pmatrix}
\cdot
\begin{pmatrix}
z \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5
\end{pmatrix}
=
\begin{pmatrix}
0 \\ L \\ W \\ B
\end{pmatrix}
$$

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 $y \in \mathbb{R}^m$.

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 condition $y^T \cdot A \geq c$ 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 (bellow) 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.*

Again, we have shifted the new basic variable from u to y and the objective function increased. The next pivot will be the circled number 3 in the third column and fourth row:

$$
\begin{array}{cccccc|c}
0 & 0 & 0 & \frac{2}{3} & \frac{5}{3} & \frac{1}{3} & \frac{31}{3} \\
\hline
0 & \boxed{1} & 0 & \frac{5}{6} & \frac{1}{3} & \frac{1}{6} & \frac{17}{3} \\
\boxed{1} & 0 & 0 & \frac{1}{6} & \frac{2}{3} & -\frac{1}{6} & \frac{4}{3} \\
0 & 0 & \boxed{1} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{array}.
$$

This is the resulting tableau, where the basic variables are $x = \frac{4}{3}, y = \frac{17}{3}, z = \frac{1}{3}$ and their values are read from the last column. Notice that all the original variables are among the basic ones and their values are non-zero. This is not always the case, see the example 3.A.1 above and its explanation via this algorithm in 3.1.6. The maximal value $\frac{31}{3}$ for the objective function is now in the right top corner.

As mentioned in the theoretical explanation, the final tableau also provides the solution of the dual problem, i.e. the minimization of $4u + 3v + 8w$ under the condition

$$
\begin{array}{rcrcrcl}
u & + & 2v & + & w & \leq & 3, \\
-u & & & + & w & \geq & 1, \\
u & + & v & + & 3w & \geq & 2.
\end{array}
$$

According to the strong duality theorem (see 3.1.5), the minimal value is again $\frac{31}{3}$, while the corresponding values of the variables u, v and w are read off the first row in the corresponding columns: $u = \frac{2}{3}, v = \frac{5}{3}, w = \frac{1}{3}$.

You may check directly that the numbers c_4, c_5, c_6 in the first row of the tableau and the value h in the top right corner satisfy $4c_4 + 3c_5 + 8c_6 = h$. Indeed, the numbers c_i tell how many times the appropriate row (the one with original value 1) has been added. Thus we obtain the right linear combination for h. □

3.A.4. Some game theory. 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. The wins and losses of such investments are well known for the last four years (for simplicity, we consider only the last four years and write them into the matrix $A = (a_{ij})$):

	gold	silver	diamonds	software
2001	2%	1%	4%	3%
2002	3%	-1%	-2%	6%
2003	1%	2%	3%	-4%
2004	-2%	1%	2%	3%

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 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 $\sum_{i=1}^{m} y_i a_{ij} > c_i$.*

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. □

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.

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) 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) 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), 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).

A very observant reader could imagine that this task is equivalent to the problem of maximizing $z_1 + z_2 + z_3 + z_4$ under the condition $A^T z \leq (1, \ldots, 1)^T$, $z \geq 0$ (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). [1]

Thus, we have to solve a linear programming problem. We introduce the slack variables w_1, w_2, w_3, w_4, and transform the problem to the standard form

$$\max \left\{ z_1 + z_2 + z_3 + z_4 \mid \left(A^T \mid E_4 \right) (z, w) = (1, 1, 1, 1)^T \right\}.$$

We work with the table:

-1	-1	-1	-1	0	0	0	0	0
2	3	1	-2	1	0	0	0	1
$\boxed{1}$	-1	2	1	0	1	0	0	1
$\textcircled{4}$	-2	3	2	0	0	1	0	1
3	6	-4	3	0	0	0	1	1

0	$-\frac{3}{2}$	$-\frac{1}{4}$	$-\frac{1}{2}$	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$
0	4	$-\frac{1}{2}$	-3	$\boxed{1}$	0	$-\frac{1}{2}$	0	$\frac{1}{2}$
0	$-\frac{1}{2}$	$\frac{5}{4}$	$\frac{1}{2}$	0	$\boxed{1}$	$-\frac{1}{4}$	0	$\frac{3}{4}$
$\boxed{1}$	$-\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$
0	$\textcircled{\frac{15}{2}}$	$-\frac{25}{4}$	$\frac{3}{2}$	0	0	$-\frac{3}{4}$	$\boxed{1}$	$\frac{1}{4}$

---

[1]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.

3.1.6. The algorithm. As already explained, the linear programming problem of maximizing the linear objective function $h = c\,x$ under the conditions $A\,x \leq b$ 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 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 line of the matrix 3.1.3(1), pick up this column and choose a line 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.

The reader should think of all the above claims in detail and check whether the algorithm must terminate. But the most striking point is the following: The slack variables parts of the matrix are closely linked to the dual linear programming problem, and there is an invariant of the entire procedure: 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 (check this!). 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 $y \cdot b$. Since the two objective functions are equal, we know that the

$$
\begin{array}{cccccccc|c}
0 & 0 & -\frac{3}{2} & -\frac{1}{5} & 0 & 0 & \frac{1}{10} & \frac{1}{5} & \frac{3}{10} \\ \hline
0 & 0 & \boxed{\frac{17}{6}} & -\frac{19}{5} & \boxed{1} & 0 & -\frac{1}{10} & -\frac{8}{15} & \frac{11}{30} \\
0 & 0 & \frac{5}{6} & \frac{3}{5} & 0 & \boxed{1} & -\frac{3}{10} & \frac{1}{15} & \frac{23}{30} \\
\boxed{1} & 0 & \frac{1}{3} & \frac{3}{5} & 0 & 0 & \frac{1}{5} & \frac{1}{15} & \frac{4}{15} \\
0 & \boxed{1} & -\frac{5}{6} & \frac{1}{5} & 0 & 0 & -\frac{1}{10} & \frac{2}{15} & \frac{1}{30}
\end{array}
$$

$$
\begin{array}{cccccccc|c}
0 & 0 & 0 & -\frac{188}{85} & \frac{9}{17} & 0 & \frac{4}{85} & -\frac{7}{85} & \frac{42}{85} \\ \hline
0 & 0 & \boxed{1} & -\frac{114}{85} & \frac{6}{17} & 0 & -\frac{3}{85} & -\frac{16}{85} & \frac{11}{85} \\
0 & 0 & 0 & \frac{146}{85} & -\frac{5}{17} & \boxed{1} & -\frac{23}{85} & \frac{19}{85} & \frac{56}{85} \\
\boxed{1} & 0 & 0 & \boxed{\frac{89}{85}} & -\frac{2}{17} & 0 & \frac{18}{85} & \frac{11}{85} & \frac{19}{85} \\
0 & \boxed{1} & 0 & -\frac{78}{85} & \frac{5}{17} & 0 & -\frac{11}{85} & -\frac{2}{85} & \frac{12}{85}
\end{array}
$$

$$
\begin{array}{cccccccc|c}
\frac{188}{89} & 0 & 0 & 0 & \frac{25}{89} & 0 & \frac{44}{89} & \frac{17}{89} & \frac{86}{89} \\ \hline
\frac{114}{89} & 0 & \boxed{1} & 0 & \frac{18}{89} & 0 & \frac{21}{89} & -\frac{2}{89} & \frac{37}{89} \\
-\frac{146}{89} & 0 & 0 & 0 & -\frac{9}{89} & \boxed{1} & -\frac{55}{89} & \frac{1}{89} & \frac{26}{89} \\
-\frac{85}{89} & 0 & 0 & \boxed{1} & -\frac{10}{89} & 0 & \frac{18}{89} & \frac{11}{89} & \frac{19}{89} \\
\frac{78}{89} & \boxed{1} & 0 & 0 & \frac{17}{89} & 0 & \frac{5}{89} & \frac{8}{89} & \frac{30}{89}
\end{array}
$$

The last table is already the optimal one, since there are no negative values in the first row. 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 $\frac{89}{86}$. $\qquad\square$

B. Difference equations

Distinct linear dependences can be an excellent tool for describing various models of growth. We begin with a very popular population model that uses a linear difference equation of second order:

3.B.1. Fibonacci sequence. In the beginning of spring, a stork brought two newborn rabbits, male and female, to a meadow. The female, after being two months old, is able to deliver two newborns, male and female. The newborns can then start delivering after one month and then every month. Every female is pregnant for one month and then she delivers. How many pairs of rabbits

algorithm provides the optimal solution. Great! (But check all the details.)

We show how all this works for the simple problem from 3.A.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:

$$
\begin{array}{cc|cccc|c}
-4 & -6 & 0 & 0 & 0 & 0 & 0 \\ \hline
1 & 2 & 1 & 0 & 0 & 0 & 120 \\
1 & 4 & 0 & 1 & 0 & 0 & 180 \\
-1 & 1 & 0 & 0 & 1 & 0 & -90 \\
1 & 0 & 0 & 0 & 0 & 1 & 110.
\end{array}
$$

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:

$$
\begin{array}{cc|cccc|c}
0 & -10 & 0 & 0 & -4 & 0 & 360 \\ \hline
0 & \boxed{3} & \boxed{1} & 0 & 1 & 0 & 30 \\
0 & 5 & 0 & \boxed{1} & 1 & 0 & 90 \\
\boxed{1} & -1 & 0 & 0 & -1 & 0 & 90 \\
0 & 1 & 0 & 0 & 1 & \boxed{1} & 20
\end{array}
$$

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 $h = 440 = 4 \cdot 90 = -4 \cdot (-90)$ 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 yilding the smalest ratio with the last right hand column entry among the positive elements – $30/3 = 10$ which is less then $90/5 = 16$ and $20/1 = 20$). This leads to the next admissible vertex in our region M and, of course the value for h will increase:

$$
\begin{array}{cc|cccc|c}
0 & 0 & \frac{10}{3} & 0 & -\frac{2}{3} & 0 & 460 \\ \hline
0 & \boxed{1} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 10 \\
0 & 0 & -\frac{5}{3} & \boxed{1} & -\frac{2}{3} & 0 & 40 \\
\boxed{1} & 0 & \frac{1}{3} & 0 & -\frac{2}{3} & 0 & 100 \\
0 & 0 & -\frac{1}{3} & 0 & \boxed{\frac{2}{3}} & \boxed{1} & 10
\end{array}
$$

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

$$
\begin{array}{cc|cccc|c}
0 & 0 & \frac{9}{3} & 0 & 0 & 1 & 470 \\ \hline
0 & \boxed{1} & \frac{1}{2} & 0 & 0 & -\frac{1}{2} & 5 \\
0 & 0 & -2 & \boxed{1} & 0 & 1 & 50 \\
\boxed{1} & 0 & 0 & 0 & 0 & 1 & 110 \\
0 & 0 & -\frac{1}{2} & 0 & \boxed{1} & \frac{3}{2} & 15
\end{array}
$$

will be there after nine months (if none of them dies and none "move in")?

Solution. After one month, there is still one pair, but the female is already pregnant. After two months, first newborns are delivered, thus there are two pairs. Every next month, there are that many new pairs as there were pregnant females one month before, which equals to the number of at least one month-old pairs, which equals the number of pairs that were there two months ago. The total number of pairs p_n after n months is thus the sum of the number of pairs in the previous two months. For the number of pairs we thus have the following *homogeneous linear recurrent formula*

$$(1) \qquad p_{n+2} = p_{n+1} + p_n, \quad n = 1, \dots,$$

which, along with the initial conditions $p_1 = 1$ and $p_2 = 1$, uniquely determines the number of pairs of rabbits at the meadow in individual months. Linearity of the formula means that all members of the sequence (p_n) appear to the first power. Hopefully the meaning of the word recurrence is clear. For the value of the n-th member we can derive an explicit formula. In searching for the formula we can use the observation 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:

$$r^{n+2} = r^{n+1} + r^n \quad \text{and after dividing by } r^n \text{ we obtain}$$
$$r^2 = r + 1.$$

This is the *characteristic equation* of the given recurrent formula. Thus our equation has roots $\frac{1-\sqrt{5}}{2}$ and $\frac{1+\sqrt{5}}{2}$ and the sequences $a_n = (\frac{1-\sqrt{5}}{2})^n$ and $b_n = (\frac{1+\sqrt{5}}{2})^n$, $n \geq 1$ satisfy the given relation. The relation is also satisfied by any linear combination, that is, any sequence $c_n = sa_n + tb_n$, $s, t \in \mathbb{R}$. The numbers s and t can be chosen so that the resulting combination satisfies the initial conditions, in our case $c_1 = 1$, $c_2 = 1$. For simplicity, it is convenient to define the zero-th member of the sequence as $c_0 = 0$ and compute s and t from the equations for c_0 and c_1. We find that $s = -\frac{1}{\sqrt{5}}$, $t = \frac{1}{\sqrt{5}}$ and thus

$$(2) \qquad p_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n(\sqrt{5})}.$$

Such a sequence satisfies the given recurrent formula and also the initial conditions $c_0 = 0$, $c_1 = 1$. Hence it is the unique sequence given by these requirements. Note that the value of p_n in the formula (2) is an integer for any natural n (all terms

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 equillibrium theorem claims we are done!

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. We can also interpret duality from this point of view. If our painter would like to quantify his efforts related to the total amount of his work by y_L per unit, the white colour painting adds y_W, while the additional work related to the black colour is y_B, then he minimizes the objective function

$$L \cdot y_L + W y_W + B y_B$$

with constraints

$$y_L + w_1 y_W + b_1 y_B \geq c_1$$
$$y_L + w_2 y_W + b_2 y_B \geq c_2.$$

But that is exactly the dual problem to the original one and the theorem 3.1.5 says that the optimal state is when the objective functions have the same value.

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.

Problems of linear programming arise with personal tasks, where workers with specific qualifications and other

in the Fibonacci sequence are integers), although it might not seem so at the first glance. □

We do some exercises about solving linear difference equation of the second order with constant coefficients. The sequence satisfying the given recurrence equation of the second order is uniquely determined whenever we prescribe any two neighbouring members. Note a further use of complex numbers: to determine the explicit formula for the n-th member of the sequence of real numbers we might require calculations with complex numbers. This happens when the characteristic polynomial of the difference equation has complex roots.

3.B.2. Find an explicit formula for the sequence satisfying the following linear difference equation with the initial conditions:

$$x_{n+2} = 2x_n + n, \;\; x_1 = 2, \; x_2 = 2.$$

Solution. The homogeneous equation is

$$x_{n+2} = 2x_n.$$

Its characteristic polynomial is $x^2 - 2$, its roots are $\pm\sqrt{2}$. The solution of the homogeneous equation is of the form

$$a(\sqrt{2})^n + b(-\sqrt{2})^n, \quad \text{for any } a, b \in \mathbb{R}.$$

We look for the particular solution using the method of indeterminate coefficients. The non-homogeneous part of the equation is a linear polynomial n. Thus a particular solution will be of the form of a linear polynomial in the variable n. That is, $kn + l$, where $k, l \in \mathbb{R}$. By substituting into the original equation we obtain

$$k(n + 2) + l = 2(kn + l) + n.$$

By comparing the coefficients of the variable n on both sides of the equation, we obtain the relation $k = 2k+1$, that is, $k = -1$. By comparing the absolute terms we obtain $2k + l = 2l$, that is, $l = -2$. Thus the particular solution is the sequence $-n - 2$.

Thus the solution of the non-homogeneous difference equation of the second order without initial condition is of the form

$$a(\sqrt{2})^n + b(-\sqrt{2})^n - n - 2, \; a, b \in \mathbb{R}$$

.

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.

3.2.1. Definition. *A homogeneous linear difference equation (or homogeneous linear recurrence) of order k* is given by the expression

$$a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k} = 0, \quad a_0 \neq 0 \quad a_k \neq 0,$$

where the coefficients a_i are scalars, which can possibly depend on n.

We usually denote the sequence in question as a function

$$x_n = f(n) = -\frac{a_1}{a_0} f(n-1) - \cdots - \frac{a_k}{a_0} f(n-k).$$

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.

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 works 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 $x_0, \ldots, x_{k-1}$ of the solution reprezent 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*

Now, by substitution in the initial conditions, we determine the indeterminate $a, b \in \mathbb{R}$. To simplify the calculation, we use a little trick: from the initial conditions and the given recurrence relation we compute the member $x_0 : x_0 = \frac{1}{2}(x_2 - 0) = 1$. The given recurrence formula along with the conditions $x_0 = 1$ and $x_1 = 1$ is then clearly satisfied by the same formula that satisfies the original initial conditions. Thus we have the following relations for a, b:

$$x_0 : \quad a(\sqrt{2})^0 + b(-\sqrt{2})^0 - 2 = 1, \quad \text{thus } a + b = 3,$$
$$x_1 : \quad \sqrt{2}a - \sqrt{2}b = 5,$$

whose solution gives us $a = \frac{6+5\sqrt{2}}{4}$, $b = \frac{6-5\sqrt{2}}{4}$. The solution is thus the sequence

$$x_n = \frac{6 + 5\sqrt{2}}{4}(\sqrt{2})^n + \frac{6 - 5\sqrt{2}}{4}(-\sqrt{2})^n - n - 2.$$

□

3.B.3. Determine the basis of the space of all solutions of the homogeneous difference equation

$$x_{n+4} = x_{n+3} + x_{n+1} - x_n,$$

Express your solution in terms of real valued functions.

Solution. The characteristic polynomial of the given equation is $x^4 - x^3 - x + 1$. If we are looking for its roots, we solve the equation

$$x^4 - x^3 - x + 1 = 0$$

The left side factors as

$$(x - 1)^2(x^2 + x + 1)$$

with two complex roots $x_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = \cos(2\pi/3) + i\sin(2\pi/3)$ and $x_2 = -\frac{1}{2} - i\frac{\sqrt{3}}{2} = \cos(2\pi/3) - i\sin(2\pi/3)$ and a double root 1. Thus the basis of the vector space of the sequences that are a solution of the difference equation in question is the following quadruple of sequences: $\{(-\frac{1}{2} + i\sqrt{3})^n\}_{n=1}^{\infty}$, $\{(-\frac{1}{2} - i\sqrt{3})^n\}_{n=1}^{\infty}$, $\{1\}_{n=1}^{\infty}$ (constant sequence) and $\{n\}_{n=1}^{\infty}$. If we are looking for a basis of real valued functions, we must replace two of the generators (sequences) from this basis by some sequences that are real only. As these generators are power series whose members are complex conjugates, it suffices to take as suitable generators the sequences given by the half of the sum and by the half of the i-th multiple of the difference of that complex generators. This yields the following real

system of solutions and all other solutions are its linear combinations.

As we have just checked, if we choose k indices i, $i + 1, \ldots, i + k - 1$ 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 $(x_n^{[1]}, \ldots, x_n^{[k]})$, it is independent if and only if the following determinant, sometimes called the *Casoratian*, is non-zero for one n

$$C_n = \begin{vmatrix} x_n^{[1]} & \cdots & x_n^{[k]} \\ x_{n+1}^{[1]} & \cdots & x_{n+1}^{[k]} \\ \vdots & \ddots & \vdots \\ x_{n+k-1}^{[1]} & \cdots & x_{n+k-1}^{[k]} \end{vmatrix} \neq 0$$

which then implies the non-vanishing of C_n for all n.

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

$$(1) \qquad f(n + 2) = a\, f(n + 1) + b\, f(n) + c,$$

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.

A similar situation with $c = 0$ and both other coefficients positive determines the famous Fibonacci sequence of numbers $y_0, y_1, \ldots$, 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 a similar problems. Thus, let us substitute into the equation (1) with coefficient $c = 0$ a similar solution as with the linear equations

basis of the solution space: $\{1\}_{n=1}^{\infty}$ (constant sequence), $\{n\}_{n=1}^{\infty}$, $\{\cos(2n\pi/3)\}_{n=1}^{\infty}$, $\{\sin(2n\pi/3)\}_{n=1}^{\infty}$. □

3.B.4. Solve the following difference equation:

$$x_{n+4} = x_{n+3} - x_{n+2} + x_{n+1} - x_n.$$

Solution. From the theory we know that the space of the solutions of this difference equation is a four-dimensional vector space whose generators can be obtained from the roots of the characteristic polynomial of the given equation. The characteristic polynomial is

$$x^4 - x^3 + x^2 - x + 1 = 0.$$

It is a reciprocal equation (that means that the coefficients at the $(n-k)$-th and k-th power of x, $k = 1, \ldots, n$, are equal). We can use the substitution $u = x + \frac{1}{x}$. After dividing the equation by x^2 (zero cannot be a root) and substituting (note that $x^2 + \frac{1}{x^2} = u^2 - 2$) we obtain

$$x^2 - x + 1 - \frac{1}{x} + \frac{1}{x^2} = u^2 - u - 1 = 0.$$

Thus we obtain the indeterminates $u_{1,2} = \frac{1 \pm \sqrt{5}}{2}$. From there then by the equation $x^2 - ux + 1 = 0$ we determine the four roots

$$x_{1,2,3,4} = \frac{1 \pm \sqrt{5} \pm \sqrt{-10 \pm 2\sqrt{5}}}{4}.$$

Note that the roots of the characteristic equation could have been "guessed" right away since

$$x^5 + 1 = (x+1)(x^4 - x^3 + x^2 - x + 1).$$

Thus the roots of the polynomial $x^4 - x^3 + x^2 - x + 1$ are also the roots of the polynomial $x^5 + 1$, which are exactly the fifth roots of the -1. By this we obtain that the solutions of the characteristic polynomial are the numbers $x_{1,2} = \cos(\frac{\pi}{5}) \pm i \sin(\frac{\pi}{5})$ and $x_{3,4} = \cos(\frac{3\pi}{5}) \pm i \sin(\frac{3\pi}{5})$. Thus the real basis of the space of the solution of the given difference equation is for instance the basis of the sequences $\cos(\frac{n\pi}{5})$, $\sin(\frac{n\pi}{5})$, $\cos(\frac{3n\pi}{5})$ and $\sin(\frac{3n\pi}{5})$, which are sines and cosines of the arguments of the corresponding powers of the roots of the characteristic polynomial.

Note that we have incidentally derived the algebraic expressions for $\cos(\frac{\pi}{5}) = \frac{1+\sqrt{5}}{4}$, $\sin(\frac{\pi}{5}) = \frac{\sqrt{10-2\sqrt{5}}}{4}$, $\cos(\frac{3\pi}{5}) = \frac{\sqrt{5}-1}{4}$ and $\sin(\frac{3\pi}{5}) = \frac{\sqrt{10+2\sqrt{5}}}{4}$. This is because all the roots of the equation have absolute value 1, they are the real and imaginary parts of the corresponding roots). □

from the first chapter (cf. 1.2.1), that is, we try $f(n) = \lambda^n$ for some scalar λ. 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 λ 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:

$$(1) \qquad y_{n+2} = y_{n+1} + \frac{1}{2}y_n, \quad y_0 = 2, \ y_1 = 0.$$

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.

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.B.5. Determine the explicit expression of the sequence satisfying the difference equation $x_{n+2} = 2x_{n+1} - 2x_n$ with initial values $x_1 = 2$, $x_2 = 2$.

Solution. The roots of the characteristic polynomial $x^2 - 2x + 2$ are $1 + i$ and $1 - i$. The basis of the (complex) vector space of the solution is thus formed by the sequences $y_n = (1+i)^n$ and $z_n = (1-i)^n$. The sequence in question can thus be expressed as a linear combination of these sequences (with complex coefficients). It is thus $x_n = a \cdot y_n + b \cdot z_n$, where $a = a_1 + ia_2$, $b = b_1 + ib_2$. From the recurrent relation we compute $x_0 = \frac{1}{2}(2x_1 - x_2) = 0$ and by substitution $n = 0$ and $n = 1$ into the expression of x_n we obtain

$$
\begin{aligned}
1 = x_0 &= a_1 + ia_2 + b_1 + ib_2 \\
2 = x_1 &= (a_1 + ia_2)(1 + i) + (b_1 + ib_2)(1 - i).
\end{aligned}
$$

By comparing the real and the complex part of both equations, we obtain a linear system of four equations with four indeterminates

$$
\begin{aligned}
a_1 + b_1 &= 1 \\
a_2 + b_2 &= 0 \\
a_1 - a_2 + b_1 + b_2 &= 2 \\
a_1 + a_2 - b_1 + b_2 &= 0.
\end{aligned}
$$

These equations imply that $a_1 = b_1 = b_2 = \frac{1}{2}$ and $a_2 = -1/2$. Thus we can express the sequence in question as

$$
x_n = \left(\frac{1}{2} - \frac{1}{2}i\right)(1+i)^n + \left(\frac{1}{2} + \frac{1}{2}i\right)(1-i)^n.
$$

The sequence can also be expressed using the real basis of the (complex) vector space of the space of solutions, that is, using the sequences $u_n = \frac{1}{2}(y_n + z_n) = (\sqrt{2})^n \cos(\frac{n\pi}{4})$ and $v_n = \frac{1}{2}i(z_n - y_n) = (\sqrt{2})^n \sin(\frac{n\pi}{4})$. The transition matrix for the changing the basis from the complex one to the real one is

$$
T := \begin{pmatrix} \frac{1}{2} & -\frac{1}{2}i \\ \frac{1}{2} & \frac{1}{2}i \end{pmatrix},
$$

the inverse matrix is $T^{-1} = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}$, for expressing the sequence x_n using the real basis, that is, for expressing the coordinates (c, d) of the sequence x_n under the basis $\{u_n, v_n\}$, we have

$$
\begin{pmatrix} c \\ d \end{pmatrix} = T^{-1} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},
$$

3.2.4. General homogeneous recurrences. We substitute $x_n = \lambda^n$ for some (yet unknown) scalar λ into the general homogeneous equation from the definition 3.2.1 (with constant coefficients). For every n we obtain the condition

$$
\lambda^{n-k}(a_0\lambda^k + a_1\lambda^{k-1} \cdots + a_k) = 0.
$$

This means that either $\lambda = 0$ or λ 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 $\lambda_1, \ldots, \lambda_k$. 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 inicial 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 $n = 0, \ldots, k - 1$ for k choices of λ_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 λ_i the determinant of such a matrix non-zero, see 2.B.7 on the page 89. 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 λ is a multiple root. We ask whether $x_n = n\lambda^n$ could be a solution. We arrive at the condition

$$
a_0 n\lambda^n + \cdots + a_k(n - k)\lambda^{n-k} = 0.
$$

This condition can be rewritten as

$$
\lambda(a_0\lambda^n + \cdots + a_k\lambda^{n-k})' = 0
$$

where the dash denotes differentiation with respect to λ (cf. the infinitesimal definition in 5.1.6, and 12.2.7 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 12.2.7 for the proof. Our condition is thus satisfied.

With greater multiplicity ℓ of the root of the characteristic polynomial we can proceed similarly and use the (now obvious) fact that a root with multiplicity ℓ is a root of all derivatives of the polynomial up to order $\ell - 1$ (inclusively). Derivatives look like this:

thus we have again an alternative expression of the sequence x_n where there are no complex numbers (but there are square roots):

$$x_n = (\sqrt{2})^n \cos\left(\frac{n\pi}{4}\right) + (\sqrt{2})^n \sin\left(\frac{n\pi}{4}\right).$$

We could have obtained these by solving two linear equations in two variables c, d, that is, $1 = x_0 = c \cdot u_0 + d \cdot v_0 = c$ and $2 = x_1 = c \cdot u_1 + d \cdot v_1 = c + d$. $\qquad\square$

3.B.6. A simplified model for the behaviour of gross domestic product. Consider the difference equation

$$(1) \qquad y_{k+2} - a(1+b)y_{k+1} + aby_k = 1,$$

where y_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.
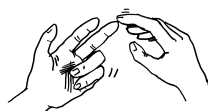
Further, we assume that the size of the domestic product is normalised such that the right-hand side of the equation is 1.

Compute the values y_n for $a = \frac{3}{4}, b = \frac{1}{3}, y_0 = 1, y_1 = 1$.

Solution. Look first for the solution of the homogeneous equation (the right side being zero) in the form of r^k. The number r must be a solution of the characteristic equation

$$x^2 - a(1+b)x + ab = 0, \quad \text{that is, } x^2 - x + \frac{1}{4} = 0,$$

which has a double root $\frac{1}{2}$. All the solutions of the homogeneous equation are then of the form $a(\frac{1}{2})^n + bn(\frac{1}{2})^n$.

Note also that if we find some solution of the non-homogeneous equation (the particular solution), then we can add to it any solution of the homogeneous solution, to obtain another solution of the non-homogeneous equation. It can be shown that all solutions of the non-homogeneous equation can be found in this way.

In this problem, it is easy to check that the constant function $y_n = c$ is a solution provided $c = 4$. All solutions of the difference equation

$$y_{k+2} - y_{k+1} + \frac{1}{4} \cdot y_k = 1$$

are thus of the form $4 + a(\frac{1}{2})^n + bn(\frac{1}{2})^n$. We require that $y_0 = y_1 = 1$ and these two equations give $a = b = -3$.

$$f(\lambda) = a_0\lambda^n + \cdots + a_k\lambda^{n-k}$$
$$f'(\lambda) = a_0 n\lambda^{n-1} + \cdots + a_k(n-k)\lambda^{n-k-1}$$
$$f''(\lambda) = a_0 n(n-1)\lambda^{n-2} + \cdots + a_k(n-k)(n-k-1)\lambda^{n-k-2}$$
$$\vdots$$
$$f^{(\ell)} = a_0 n\cdots(n-\ell+1)\lambda^{n-\ell} + \cdots$$
$$\qquad + a_k(n-k)\cdots(n-k-\ell+1)\lambda^{n-k-\ell}.$$

We look at the case of a triple root λ and try to find a solution in the form $n^2\lambda^n$. By substitution into the definition, we obtain the equation

$$a_0 n^2\lambda^n + \cdots + a_k(n-k)^2\lambda^{n-k} = 0.$$

Clearly the left side equals the expression $\lambda^2 f''(\lambda) + \lambda f'(\lambda)$ and because λ 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 $x_n = n^\ell\lambda^n$,

$$a_0 n^\ell\lambda^n + \ldots a_k(n-k)^\ell\lambda^{n-k} = 0,$$

the solution can be obtained as a linear combination of the derivatives of the characteristic polynomial starting with the expression (check the combinatorics!)

$$\lambda^\ell f^{(\ell)} + \binom{\ell}{2}\lambda^{\ell-1} f^{(\ell-1)} + \ldots.$$

We have thus come close to the complete proof of the following:

HOMOGENOUS EQUATIONS WITH CONSTANT COEFFICIENTS

Theorem. *The solution space of a homogeneous linear difference equation of order k 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 λ are (complex) roots of the characteristic polynomial and the powers ℓ run over all natural numbers $0, \ldots, r_\lambda - 1$, where r_λ is the multiplicity of the root λ.*

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.

To illustrate of our approach we show how the calculation looks for the case of a root λ_1 with multiplicity one and a root

Thus the solution of this non-homogeneous equation is

$$y_n = 4 - 3\left(\frac{1}{2}\right)^n - 3n\left(\frac{1}{2}\right)^n.$$

Again, as we know that the sequence given by this formula satisfies the given difference equation and also the given initial conditions, it is indeed the only sequence characterized by these properties.

□

3.B.7. Find a sequence which satisfies the given non-homogeneous difference equation with the initial conditions:

$$x_{n+2} = x_{n+1} + 2x_n + 1, \quad x_1 = 2,\, x_2 = 2.$$

Solution. The general solution of the homogeneous equation is of the form $a(-1)^n + b2^n$. A particular solution is the constant $-1/2$. The general solution of the given non-homogeneous equation without initial conditions is thus

$$a(-1)^n + b2^n - \frac{1}{2}.$$

Substituting in the initial conditions, then gives the constants $a = -5/6$, $b = 5/6$. The given difference equation with initial conditions is thus satisfied by the sequence

$$x_n = -\frac{5}{6}(-1)^n + \frac{5}{3}2^{n-1} - \frac{1}{2}.$$

□

3.B.8. Determine the sequence of real numbers that satisfies the following non-homogeneous difference equation with initial conditions:

$$2x_{n+2} = -x_{n+1} + x_n + 2, \quad x_1 = 2,\, x_2 = 3.$$

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

$$a(-1)^n + b\left(\frac{1}{2}\right)^n + 1.$$

By substitution with the initial conditions, we obtain the constants $a = 1$, $b = 4$. The given equation with initial conditions is thus satisfied by the sequence

$$x_n = (-1)^n + 4\left(\frac{1}{2}\right)^n + 1.$$

□

λ_2 with multiplicity two:

$$C(\lambda_1^n, \lambda_2^n, n\lambda_2^n) = \begin{vmatrix} \lambda_1^n & \lambda_2^n & n\lambda_2^n \\ \lambda_1^{n+1} & \lambda_2^{n+1} & (n+1)\lambda_2^{n+1} \\ \lambda_1^{n+2} & \lambda_2^{n+2} & (n+2)\lambda_2^{n+2} \end{vmatrix}$$

$$= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 & \lambda_2 & (n+1)\lambda_2 \\ \lambda_1^2 & \lambda_2^2 & (n+2)\lambda_2^2 \end{vmatrix}$$

$$= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 - \lambda_2 & 0 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & 0 & \lambda_2^2 \end{vmatrix}$$

$$= -\lambda_1^n \lambda_2^{2n} \begin{vmatrix} \lambda_1 - \lambda_2 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & \lambda_2^2 \end{vmatrix}$$

$$= \lambda_1^n \lambda_2^{2n+1}(\lambda_1 - \lambda_2)^2 \neq 0.$$

In the general case the proof can be carried on inductively in a similar way. □

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

3.B.9. Determine sequences satisfying

$$x_{n+2} - 6x_{n+1} + 5x_n = n\,\mathrm{e}^n\,.$$

Solution. Solve first the homogeneous part. We get:

$$x_n^{(h)} = c_1 \cdot (1)^n + c_2 \cdot 5^n.$$

To find the particular solution we can use the method of variation of the constant. The Wronski determinant is

$$W_{j+1} = \det \begin{pmatrix} 1^{j+1} & 5^{j+1} \\ 1^{j+2} & 5^{j+2} \end{pmatrix} = 4 \cdot 5^{j+1}.$$

Thus,

$$x_n = c_1 + c_2 \cdot 5^n - \frac{1}{4}\sum_{j=0}^{n-1} j\,\mathrm{e}^j + \left(\sum_{j=0}^{n-1} \frac{j\,\mathrm{e}^j}{4 \cdot 5^{j+1}}\right) 5^n,$$

$c_1, c_2 \in \mathbb{R}$. □

3.B.10. 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 $x_2 = 3$. ○

3.B.11. Determine an explicit formula for the n-th member of the unique solution $\{x_n\}_{n=1}^\infty$ that satisfies the following conditions:

$$x_{n+2} = x_{n+1} - x_n,\ x_1 = 1,\ x_2 = 5.$$

○

3.B.12. Determine an explicit formula for the n-th member of the unique solution $\{x_n\}_{n=1}^\infty$ that satisfies the following conditions:

$$-x_{n+3} = 2x_{n+2} + 2x_{n+1} + x_n,\ x_1 = 1,\ x_2 = 1,\ x_3 = 1.$$

○

3.B.13. Determine an explicit formula for the n-th member of the unique solution $\{x_n\}_{n=1}^\infty$ that satisfies the following conditions:

$$-x_{n+3} = 3x_{n+2} + 3x_{n+1} + x_n,\ x_1 = 1,\ x_2 = 1,\ x_3 = 1.$$

○

C. Population models

Population models, which we consider now, have recurrence relations in vector spaces. The unknown in this case is not a sequence of numbers but a sequence of vectors. The role of coefficients is played by matrices. We begin with a simple (two-dimensional) case.

difference of two solutions of a non-homogeneous system is 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

$$x_n = \alpha_0 + \alpha_1 n + \cdots + \alpha_s n^s$$

with unknown coefficients α_i, $i = 1, \ldots, s$. 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 α_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 α_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.B.6.

3.2.7. Variation of constants. An other 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

$$x_n = A_n x_n^{(1)} + B_n x_n^{(1)}$$

3.C.1. Savings. A friend and I save for a holiday together by monthly payments in the following way. At the beginning I give 10 €and he gives 20 €. Every consecutive month each of us gives as many as last month plus one half of what the other has given the month before. How much will we have after one year? How much money will I pay in the twelfth month?

Solution. Let the amount of money I pay in the n-th month be denoted by x_n, and the amount my friend pays is y_n. Thus in the first month we deposit $x_1 = 10, y_1 = 20$. For the following payments we can write down a recurrent relation:

$$x_{n+1} = x_n + \tfrac{1}{2}y_n$$
$$y_{n+1} = y_n + \tfrac{1}{2}x_n$$

If we denote the common savings by $z_n = x_n + y_n$, then by summing the equations we obtain $z_{n+1} = z_n + \tfrac{1}{2}z_n = \tfrac{3}{2}z_n$. This is a geometric sequence and we obtain $z_n = 3.(\tfrac{3}{2})^{n-1}$. In a year we will have $z_1 + z_2 + \cdots + z_{12}$. This partial sum is easy to compute

$$3\left(1 + \frac{3}{2} + \cdots + \left(\frac{3}{2}\right)^{11}\right) = 3\frac{(\frac{3}{2})^{12} - 1}{\frac{3}{2} - 1} \doteq 772,5.$$

In a year we will have saved over 772 €.

The recurrent system of equation describing the savings system can be written by matrices as follows:

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

It is thus again a geometric sequence. Its elements are now vectors and the quotient is not a scalar, but a matrix. The solution can be found analogously:

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}^{n-1} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$

The power of the matrix acting on the vector (x_1, y_1) can be found by expressing this vector in the basis of eigenvectors. The characteristic polynomial of the matrix is $(1-\lambda)^2 - \tfrac{1}{4} - 0$ and the eigenvalues are thus $\lambda_{1,2} = \tfrac{3}{2}, \tfrac{1}{2}$. The corresponding eigenvectors are thus $(1, 1)$ and $(1, -1)$. For the initial vector $(x_1, y_1) = (1, 2)$ we compute

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \frac{3}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and thus

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \frac{3}{2}\left(\frac{3}{2}\right)^{n-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}\left(\frac{1}{2}\right)^{n-1}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

That means that in the 12th month I pay

$$x_{12} = \left(\frac{3}{2}\right)^{12} - \left(\frac{1}{2}\right)^{12} \doteq 130$$

with some conditions on A_n and B_n to be imposed. We have

$$x_{n+1} = A_{n+1}x_{n+1}^{(1)} + B_{n+1}x_{n+1}^{(2)} = A_n x_{n+1}^{(1)} + B_n x_{n+1}^{(2)} +$$
$$(A_{n+1} - A_n)x_{n+1}^{(1)} + (B_{n+1} - B_n)x_{n+1}^{(2)}$$
$$= A_n x_{n+1}^{(1)} + B_n x_{n+1}^{(2)} + \delta A_n x_{n+1}^{(1)} + \delta B_n x_{n+1}^{(2)},$$

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

$$\delta A_n x_{n+1}^{(1)} + \delta B_n x_{n+1}^{(2)} = 0.$$

Thus, for all n

$$x_{n+1} = A_n x_{n+1}^{(1)} + B_n x_{n+1}^{(2)},$$

and in particular

$$x_{n+2} = A_{n+1}x_{n+2}^{(1)} + B_{n+1}x_{n+2}^{(2)}$$
$$= A_n x_{n+2}^{(1)} + B_n x_{n+2}^{(2)} + \delta A_n x_{n+2}^{(1)} + \delta B_n x_{n+2}^{(2)}.$$

Now,

$$f_n = x_{n+2} + a_n x_{n+1} + b_n x_{n+2}$$
$$= A_n\left(x_{n+2}^{(1)} + a_n x_{n+1}^{(1)} + b_n x_{n+2}^{(1)}\right) + B_n\left(x_{n+2}^{(2)} + \right.$$
$$\left. a_n x_{n+1}^{(2)} + b_n x_{n+2}^{(2)}\right) + \delta A_n x_{n+2}^{(1)} + \delta B_n x_{n+2}^{(2)}$$
$$= \delta A_n x_{n+2}^{(1)} + \delta B_n x_{n+2}^{(2)}$$

Hence the variations δA_n and δB_n are subject to the systems

$$\delta A_n x_{n+1}^{(1)} + \delta B_n x_{n+1}^{(2)} = 0$$
$$\delta A_n x_{n+2}^{(1)} + \delta B_n x_{n+2}^{(2)} = f_n$$

with solutions (compute the inverse matrix e.g. by means of the algebraic adjoint and the determinant)

$$\delta A_n = A_{n+1} - A_n = -\frac{f_n x_{n+1}^{(2)}}{W_{n+1}}$$

$$\delta B_n = B_{n+1} - B_n = \frac{f_n x_{n+1}^{(1)}}{W_{n+1}}$$

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

$$A_n - A_0 = \sum_{j=0}^{n-1} -\frac{f_j x_{j+1}^{(2)}}{W_{j+1}}$$

$$B_n - B_0 = \sum_{j=0}^{n-1} \frac{f_j x_{j+1}^{(1)}}{W_{j+1}}.$$

€ and my friend pays basically the same amount. □

Remark. The previous example can be solved also without matrices by rewriting the recurrent equation: $x_{n+1} = x_n + \frac{1}{2}y_n = \frac{1}{2}x_n + \frac{1}{2}z_n$.

The previous example was actually a model of growth (in this case, growth of saved money). We go now to the models of growth describing primarily the growth of a population. The Leslie model of population growth with which we have dealt with in great detail in the theoretical part describes very well not only populations of sheep (according to which it was developed), but can be also applied in modelling of the following populations:

3.C.2. Rabbits for the second time. We show how the Leslie model can describe the population of the rabbits on a meadow with which we have worked in exercise (3.B.1). Suppose that the rabbits are dying after reaching the ninth year of age (in the original model the rabbits were immortal). Denote the number of rabbits according to their age in months at time t (in months) as $x_1(t)$, $x_2(t)$,..., $x_9(t)$. Then the number of rabbits in individual categories are described after one month by the formula $x_1(t+1) = x_2(t)+x_3(t)+\cdots+x_9(t)$, $x_i(t+1) = x_{i-1}(t)$, for $i = 2, 3, \ldots, 10$, or

$$
\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ x_4(t+1) \\ x_5(t+1) \\ x_6(t+1) \\ x_7(t+1) \\ x_8(t+1) \\ x_9(t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \\ x_7(t) \\ x_8(t) \\ x_9(t) \end{pmatrix}.
$$

The characteristic polynomial of the given matrix is $\lambda^9 - \lambda^7 - \lambda^6 - \lambda^5 - \lambda^4 - \lambda^3 - \lambda^2 - \lambda - 1$. The roots of this polynomial are hard to explicitly express, but we can estimate one of them very well – $\lambda_1 \doteq 1.608$ (why must it be smaller than $(\sqrt{5} + 1)/2)$?). Thus the population grows according to this model approximately with the geometric sequence 1.608^t.

3.C.3. Pond. Suppose we have a simple model of a pond where there lives a population of white fish (roach, bleak, vimba, nase, etc.). Assume that 20 % of babies survive their second year and from that age on they are able to reproduce. For these young fish, approximately 60 % of them survive their third year and in the following years the mortality can be ignored. Furthermore we assume that the birth rate is three times the number of fish that can reproduce.

Such a population would clearly fill the pond very quickly. Thus we want to maintain a balance by using a

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 aquired 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.B.9.

3.2.8. Linear filters. Now we consider infinite sequences

$$
x = (\ldots, x_{-n}, x_{-n+1}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n, \ldots).
$$

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

$$
z_n = a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k}.
$$

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

$$
x_{n+p} = x_n
$$

will also have periodic images $z = Tx$

$$
z_{n+p} = a_0 x_{n+p} + a_1 x_{n-1+p} + \cdots + a_k x_{n-k+p}
$$
$$
= a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k} = z_n
$$

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. Also, we are looking for the kernel of our linear mapping T. That is, the subspace of sequences given by the homogeneous difference equation

$$
a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k} = 0, \quad a_0 \neq 0 \quad a_k \neq 0,
$$

which we are able to solve.

157

predator, for instance esox. Assume that one esox eats per year approximately 500 mature white fish. How many esox should be put into the pond in order for the population to remain constant?

Solution. If we denote by p the number of babies, by m the number of young fish and by r the number of adult fish, then the state of the population in the next year is given by:

$$\begin{pmatrix} p \\ m \\ r \end{pmatrix} \mapsto \begin{pmatrix} 3m + 3r \\ 0.2p \\ 0.6m + \tau r \end{pmatrix},$$

where $1 - \tau$ is the relative mortality of the adult fish caused by the esox. The corresponding matrix describing this model is then

$$\begin{pmatrix} 0 & 3 & 3 \\ 0.2 & 0 & 0 \\ 0 & 0.6 & \tau \end{pmatrix}$$

If the population is to stagnate, (ie. remain constant), then this matrix must have eigenvalue 1. In other words, one must be the root of the characteristic polynomial of this matrix. That is of the form $\lambda^2(\tau - \lambda) + 0.36 - 0.6.(\tau - \lambda) = 0$. That means that τ must satisfy

$$\tau - 1 + 0.36 - 0.6(\tau - 1) = 0$$
$$0.4\tau - 0.04 = 0$$

In the next year only 10 % is allowed to survive and the rest should be eaten by the esox. If we denote the desired number of esox by x, then together they eat $500x$ fish, which, according to the previous computation, should be $0.9r$. The ratio of the number of white fish to the number of esox should thus be $\frac{r}{x} = \frac{500}{0.9}$. That is, one esox for (approximately) 556 white fish. $\qquad\square$

3.C.4. In the population model, let the number of predators be D_k and the number of preys be K_k in month k. The relation of these between month k and month $k + 1$ is given by one of the three linear systems

(a)

$$\begin{aligned} D_{k+1} &= & 0.6\,D_k &+ & 0.5\,K_k, \\ K_{k+1} &= & -0.16\,D_k &+ & 1.2\,K_k; \end{aligned}$$

(b)

$$\begin{aligned} D_{k+1} &= & 0.6\,D_k &+ & 0.5\,K_k, \\ K_{k+1} &= & -0.175\,D_k &+ & 1.2\,K_k; \end{aligned}$$

(c)

$$\begin{aligned} D_{k+1} &= & 0.6\,D_k &+ & 0.5\,K_k, \\ K_{k+1} &= & -0.135\,D_k &+ & 1.2\,K_k. \end{aligned}$$

Analyse the behaviour of this model for large time values.

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.
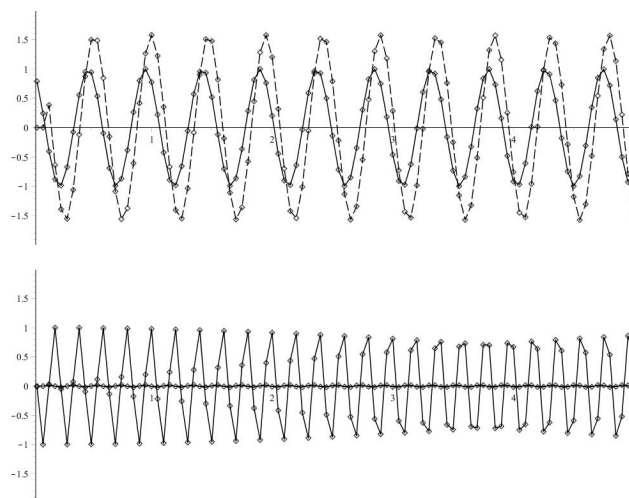
The first case shows an amplifying of the signal, while the second frequency is close to the kernel which is killed by the filter. Notice that the filtered signal suffers serious shifts in phase, which varies with the frequencies. Cheap equalisers work in such a bad way.

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 equation require 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 intital conditions can be described as an iterated process. Imagine we keep the state vector of the

Solution. Note that the individual variants differ from each other only in the value of the coefficients at D_k in the second equation. Thus we can express all three cases as

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} = \begin{pmatrix} 0.6 & 0.5 \\ -a & 1.2 \end{pmatrix} \cdot \begin{pmatrix} D_{k-1} \\ K_{k-1} \end{pmatrix}, \quad k \in \mathbb{N},$$

where we set $a = 0.16$, $a = 0.175$, $a = 0.135$. The value of the coefficient a represents here the average number of preys killed by one predator per month. When denoting

$$T = \begin{pmatrix} 0.6 & 0.5 \\ -a & 1.2 \end{pmatrix}$$

we obtain

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} = T^k \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix}, \quad k \in \mathbb{N}.$$

Using the powers of the matrix T we can determine the evolution of the populations of predators and prey after a very long time.

We compute the eigenvalues for the matrix T

(a) $\lambda_1 = 1$, $\quad \lambda_2 = 0.8$;

(b) $\lambda_1 = 0.95$, $\quad \lambda_2 = 0.85$;

(c) $\lambda_1 = 1.05$, $\quad \lambda_2 = 0.75$.

The respective eigenvectors are

(a) $(5, 4)^T$, $\quad (5, 2)^T$;

(b) $(10, 7)^T$, $\quad (2, 1)^T$;

(c) $(10, 9)^T$, $\quad (10, 3)^T$.

For $k \in \mathbb{N}$ we obtain

(a)

$$T^k = \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \end{pmatrix}^k \cdot \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix}^{-1};$$

(b)

$$T^k = \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.95 & 0 \\ 0 & 0.85 \end{pmatrix}^k \cdot \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix}^{-1};$$

(c)

$$T^k = \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1.05 & 0 \\ 0 & 0.75 \end{pmatrix}^k \cdot \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix}^{-1}.$$

From there we have for large $k \in \mathbb{N}$ that

(a)

$$T^k \approx \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix}^{-1}$$
$$= \frac{1}{10} \begin{pmatrix} -10 & 25 \\ -8 & 20 \end{pmatrix};$$

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 $x_{n+1} = a_1 x_n + \cdots + a_k x_{n-k+1}$ 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:

$$A = \begin{pmatrix} a_1 & a_2 & \dots & a_{k-1} & a_k \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

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

$$p(\lambda) = (-1)^k (\lambda^k - a_1 \lambda^{k-1} - \cdots - a_k),$$

as we can check directly or by expanding the last column and employing induction on k.

Thus, the eigenvalues are exactly the roots λ 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 λ-multiple. Thus every such λ 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}.

(b)

$$T^k \approx \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix}^{-1}$$
$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};$$

(c)

$$T^k \approx \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1.05^k & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix}^{-1}$$
$$= \frac{1.05^k}{60} \begin{pmatrix} -30 & 100 \\ -27 & 90 \end{pmatrix},$$

because for large $k \in \mathbb{N}$ we can set

(a)

$$\begin{pmatrix} 1 & 0 \\ 0 & 0.8 \end{pmatrix}^k \approx \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix};$$

(b)

$$\begin{pmatrix} 0.95 & 0 \\ 0 & 0.85 \end{pmatrix}^k \approx \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};$$

(c)

$$\begin{pmatrix} 1.05 & 0 \\ 0 & 0.75 \end{pmatrix}^k \approx \begin{pmatrix} 1.05^k & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that in variant (b), that is for $a = 0.175$, it is not necessary to compute the eigenvectors.

Thus we have

(a)

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} \approx \frac{1}{10} \begin{pmatrix} -10 & 25 \\ -8 & 20 \end{pmatrix} \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix}$$
$$= \frac{1}{10} \begin{pmatrix} 5\,(-2D_0 + 5K_0) \\ 4\,(-2D_0 + 5K_0) \end{pmatrix};$$

(b)

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} \approx \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

(c)

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} \approx \frac{1.05^k}{60} \begin{pmatrix} -30 & 100 \\ -27 & 90 \end{pmatrix} \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix}$$
$$= \frac{1.05^k}{60} \begin{pmatrix} 10\,(-3D_0 + 10K_0) \\ 9\,(-3D_0 + 10K_0) \end{pmatrix}.$$

These results can be interpreted as follows:

(a) If $2D_0 < 5K_0$, the sizes of both populations stabilise on non-zero sizes (we say that they are stable); if $2D_0 \geq 5K_0$, both populations die out.

(b) Both populations die out.

(c) For $3D_0 < 10K_0$ begins a population boom of both kinds; for $3D_0 \geq 10K_0$ both populations die out.

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 τ_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 τ_i are between zero and one. Note that when all τ are equal one, it is actually a linear recurrence with constant coefficients and thus has either exponential growth/decay (for real roots λ 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 τ_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. 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 λ positive but very small the value of q will be arbitrarily large, while for large λ, it will be arbitrarily close to zero. Thus, evidently there exists exactly one positive λ for which $q(\lambda) = 1$ and thus also $p_m(\lambda) = 0$. In other words, for every Leslie matrix (with all the parameters f_i and τ_i positive), there exists exactly one positive real eigenvalue. For actual Leslie models of populations a typical situation is when the only real eigenvalue λ_1 is greater or equal to one, while the absolute values of the other eigenvalues are strictly less than one.

If we begin with any state vector X, given as a sum of eigenvectors

$$X = X_1 + \dots + X_m$$

Even a small change in the size of a can lead to a completely different result. This is caused by the constancy of the value of a: it does not depend on the size of the populations. Note that this restriction (that is, assuming a to be constant) has no interpretation in reality. But still we obtain an estimate on the sizes of a for stable populations. □

3.C.5. Remark. Another model for the populations of predators and preys is the model by Lotka and Volterra, which describes a relation between the populations by a system of two ordinary differential equations. Using this model both populations oscillate, which is in accord with observations.

Other interesting and well-described models of growth can be found in the collection of exercises after this chapter. (see 3.G.2

In linear models an important role is played by primitive matrices (3.3.3).

3.C.6. Which of the matrices

$$A = \begin{pmatrix} 0 & 1/7 \\ 1 & 6/7 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 1/6 \end{pmatrix},$$

$$C = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 0 & 1/2 \\ 3/4 & 0 & 1/2 \end{pmatrix}, \quad D = \begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/6 & 1/6 & 1/3 \\ 1/6 & 0 & 5/6 & 2/3 \end{pmatrix}$$

$$E = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

are primitive?

Solution.

$$A^2 = \begin{pmatrix} 1/7 & 6/49 \\ 6/7 & 43/49 \end{pmatrix}, \quad C^3 = \begin{pmatrix} 3/8 & 1/4 & 1/4 \\ 1/4 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/2 \end{pmatrix}.$$

So the matrices A and C are primitive, since (respectively) A^2 and C^3 are positive matrices. The middle column of the matrix B^n is always (for $n \in \mathbb{N}$) the vector $(0, 1, 0)^T$ which contains the entry 0. Hence the matrix B cannot be primitive. The product

$$\begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/6 & 1/6 & 1/3 \\ 1/6 & 0 & 5/6 & 2/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ a/6 + b/3 \\ 5a/6 + 2b/3 \end{pmatrix}, a, b \in \mathbb{R}$$

$a, b \in \mathbb{R}$, 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 $D^3 = D \cdot D^2$,

with eigenvalues λ_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 $\lambda_1 X_1^k$.

The distribution of the population among the age groups are thus very fast approaching the ratios of the components of the eigenvector to the dominant eigenvalue λ_1.

As an example, consider the matrix below where individual coefficients are taken from the model for sheep breeding, that is, the values τ 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 τ_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.[2]

We begin with some definitions in order to formulate our ideas.

_____

[2] 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.

$D^4 = D \cdot D^3, \ldots, D^n = D \cdot D^{n-1}, \ldots$, thus the matrix D is not primitive. The matrix E is a permutation matrix (in every row and every column there is exactly one non-zero element, 1). It is not difficult to see that a power of a permutation matrix is again a permutation matrix. Thus the matrix E is also not primitive. This is easily verified by calculating the powers E^2, E^3, E^4. The matrix E^4 is a unit matrix. $\square$

D. Markov processes

3.D.1. Sweet-toothed gambler. A gambler bets on a coin – whether a flip results in a head or in a tail. At the start of the game he has three sweets. On every flip, he bets on a sweet. If he wins, he gains one additional sweet. If he looses, he looses the sweet. The game ends when he loses all sweets or has at least five sweets. What is the probability that the game does not end after four bets?

Solution. Before the j-th round we can describe the state of the player by the random vector $X_j = (p_0(j), p_1(j), p_2(j), p_3(j), p_4(j), p_5(j))$, where p_i is the probability that the player has i sweets. If the player has before the j-th bet i sweets ($i = 2, 3, 4$), then after the bet he has $(i - 1)$ sweets with probability $1/2$, and he has $(i + 1)$ sweets with probability $1/2$. If he attains five sweets or loses them all, the number of sweets does not change. The vector X_{j+1} is then obtained from the vector X_j by multiplying it with the matrix

$$
A := \begin{pmatrix}
1 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.5 & 1
\end{pmatrix}.
$$

At the start,

$$
X_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.
$$

After four bets the situation is described by the vector

$$
X_5 = A^4 X_1 = \begin{pmatrix} 1/8 \\ 3/16 \\ 0 \\ 5/16 \\ 0 \\ 3/8 \end{pmatrix},
$$

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.

Theorem. *If A is a primitive matrix with spectral radius $\lambda \in \mathbb{R}$, then λ is a root of the characteristic polynomial of A with multiplicity one and λ is strictly greater than the absolute value of all other eigenvalues of A. Furthermore, there exists an eigenvector x associated with λ such that all elements x_i of x are positive.*

PROOF. We shall present rather a sketch of the proof and we shall rely on intuition from elementary geometry.

Notice that the matrices A and A^k share the eigenvectors, while the corresponding eigenvalues are λ and λ^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

$$
S = \{ x = (x_1, \ldots, x_n)^T, |x| = 1, x_i \geq 0, i = 1, \ldots, n \}.
$$

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 $S \to S$ satisfies all the assumptions of the *Brouwer fixed point theorem*[3] and thus there

---

[3]This theorem is a great example of a blend of (homological) Algebra, (differential) Topology and Analysis. We shall discuss it in Chapter 9, cf. **??** on page **??**.

that is, the probability that the game ends in the fourth bet or sooner is one half.

Note that the matrix A describing the evolution of the probabilist vector X is itself probabilistic, that is, in each column the sum is one. But it does not have the property required by the Perron-Frobenius theorem. By a simple computation you can check (or you can see it straight without any computation) that there exist two linearly independent eigenvectors corresponding to the eigenvalue 1. These correspond to the case that the player has no sweet, that is $x = (1, 0, 0, 0, 0, 0)^T$, or to the case when the player has 5 sweets and the game thus ends with him keeping all the sweets, that is, $x = (0, 0, 0, 0, 0, 1)^T$. All other eigenvalues (approximately 0.8, 0.3, -0.8, -0.3) are in absolute value strictly smaller than one. Thus the components in the corresponding eigensubspaces with iteration of the process with arbitrary initial distribution vanish and the process approaches the limiting value of the probabilistic vector of the form $(a, 0, 0, 0, 0, 1 - a)$, where the value a depends on the initial number of sweets. In our case it is $a = 0.4$, if there were 4 sweets at the start, it would be $a = 0.2$ and so on. $\square$

3.D.2. Car rental. A company that rents cars every week has two branches – one in Prague and one in Brno. A car rented in Brno can be returned in Prague and vice versa. After some time it has been discovered that in Prague, roughly 80 % of the cars rented in Prague and 90 % of the cars rented in Brno are returned there. How to distribute the cars among the branches such that in both there is at the start of the week always the same number of cars as in the week before? How will the situation look like after a long time, if the cars are distributed at the start in a random way?

Solution. Denote the components of the vector in question, that is, the initial number of cars in Brno and in Prague by x_B and x_P respectively. The distribution of the cars between branches is then described by the vector $x = \begin{pmatrix} x_B \\ x_P \end{pmatrix}$. If we consider such a multiple of the vector x such that the sum of its components in 1, then its components give the percentage distribution of the cars. According to the statement, the state at the end of the week is described by the vector $\begin{pmatrix} 0.1 & 0.2 \\ 0.9 & 0.8 \end{pmatrix} \begin{pmatrix} x_B \\ x_P \end{pmatrix}$. The matrix $A = \begin{pmatrix} 0.1 & 0.2 \\ 0.9 & 0.8 \end{pmatrix}$ thus describes our (linear) system of car rental. If at the end of the week in the branches there should be the same number of cars as at the beginning, we are looking for such a vector x for

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 $(\lambda, \ldots, \lambda)$. Moreover, We multiply the mapping by by the constant λ^{-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 $z = (1, \ldots, 1)^T$ 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 thoerem is finished.

In order to do that we use an auxiliary lemma. Consider for the moment the matrix B to define the linear mapping that maps the row vectors

$$u = (u_1, \ldots, u_n) \mapsto u \cdot B = v,$$

that is, using multiplication from the right (i.e. B is viewed as the matrix of a linear map on one-forms). Since $z = (1, \ldots, 1)^T$ is an eigenvector of the matrix B (with eigenvalue 1), the sum of the coordinates of the row vector v

$$u \cdot B \cdot (1, \ldots, 1)^T = \sum_{i,j=1}^{n} u_i b_{ij} = \sum_{i=1}^{n} u_i = 1,$$

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 is positive by our assumption, the image of the simplex S under B 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. Moreover, 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. Restriction of our mapping to P, and P itself satisfy the assumptions of the auxiliary lemma proved 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.

The theorem is nearly proved. We have just to consider the problem that the mapping under question was given by

which that $Ax = x$. That means that we are looking for an eigenvector of the matrix A associated with the eigenvalue 1.

The characteristic polynomial of the matrix A is $(0.1 - \lambda)(0.8 - \lambda) - 0.9 \cdot 0.2 = (\lambda - 1)(\lambda + 0.1)$ and 1 is indeed an eigenvalue of the matrix A. The corresponding eigenvector $x = \begin{pmatrix} x_B \\ x_P \end{pmatrix}$ satisfies the equation $\begin{pmatrix} -0.9 & 0.2 \\ 0.9 & -0.2 \end{pmatrix} \begin{pmatrix} x_B \\ x_P \end{pmatrix} = 0$. It is thus a multiple of the vector $\begin{pmatrix} 0.2 \\ 0.9 \end{pmatrix}$. For determining the percentage distribution we are looking for a multiple such that $x_B + x_P = 1$. That is satisfied by the vector $\frac{1}{1,1} \begin{pmatrix} 0.2 \\ 0.9 \end{pmatrix} = \begin{pmatrix} 0.18 \\ 0.82 \end{pmatrix}$. The suitable distribution of the cars between Prague and Brno is such that 18% of the cars are in Brno and 82% of the cars are in Prague.

If we choose arbitrarily the initial state $x = \begin{pmatrix} x_B \\ x_P \end{pmatrix}$, then the state after n weeks is described by the vector $x_n = A^n x$. It is useful to express the initial vector x in the basis of the eigenvectors of A. The eigenvector for the eigenvalue 1 has already been found. Similarly we find eigenvectors for the eigenvalue -0.1. That is for instance the vector $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

The initial vector can be expressed as a linear combination $x = a \begin{pmatrix} 0.2 \\ 0.9 \end{pmatrix} + b \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. The state after n weeks is then

$$x_n = A^n (a \begin{pmatrix} 0.18 \\ 0.82 \end{pmatrix} + b \begin{pmatrix} -1 \\ 1 \end{pmatrix})$$
$$= a \begin{pmatrix} 0.18 \\ 0.82 \end{pmatrix} + b(-0.1)^n \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

The second summand is approaching zero for $n \to \infty$. Thus the state stabilises at $a \begin{pmatrix} 0.18 \\ 0.82 \end{pmatrix}$. That is, the coordinate of the initial vector at the direction of the first eigenvector. The coefficient can be easily expressed using the initial states of the cars: $a = \frac{x_B + x_P}{1,1}$. $\qquad \square$

3.D.3. 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.

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 proven the claim about eigenvalues for the transpose of B. But transposing does not change the eigenvalues and so the proof is complete. $\qquad \square$

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.

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 $\psi : \mathbb{R}^n \to \mathbb{R}^n$ 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 ψ is strictly less than one.*

PROOF. Consider the matrix A of the mapping ψ 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 ψ already maps P into P. Clearly ψ cannot have any eigenvalue with absolute value greater than one.

We argue by contradiction and assume that there exists an eigenvalue λ with $|\lambda| = 1$. Then there are two possibilities, either $\lambda^k = 1$ for suitable 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 ψ 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 λ is one, because such a fixed point of ψ^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 ψ 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. $\qquad \square$

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.

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 for the 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_∞, 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 the player A or player B. Thus we can expect that 60 % of the games will be played against the worse of the two opponents. Because

$$\frac{2}{5} = \frac{3}{5} \cdot \frac{1}{2} + \frac{2}{5} \cdot \frac{1}{4},$$

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

$$\begin{pmatrix} 3/4 & 9/16 \\ 1/4 & 7/16 \end{pmatrix}.$$

It is easily determined that now

$$x_\infty = \left(\frac{9}{13}, \frac{4}{13}\right)^T.$$

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. The second strategy is thus better. Note also that for the second strategy, roughly 42,3 % of the games are winning ones. It suffices to enumerate

$$0.423 \doteq \frac{11}{26} = \frac{9}{13} \cdot \frac{1}{2} + \frac{4}{13} \cdot \frac{1}{4}.$$

$\square$

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 λ, 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 μ 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 (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:

$$\sum_{i,j=1}^{n} a_{ij}x_j = \sum_{i=1}^{n} \lambda x_i = \lambda$$

$$\lambda = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}\right) x_j \leq \sum_{j=1}^{n} b_{\max}x_j = b_{\max}$$

$$\lambda = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}\right) x_j \geq \sum_{j=1}^{n} b_{\min}x_j = b_{\min}. \qquad \square$$

Note that for instance all Leslie matrices from 3.3.2, as soon as all their parameters f_i and τ_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 form 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*) $x_n = (u_1(n), \dots, u_m(n))^T$.

This means that all components of the vector x are real non-negative numbers and their sum equals one. Components

3.D.4. Absent-minded professor. Consider the following situation. An absent-minded professor carries an umbrella with him, but with probability $1/2$ he forgets it from wherever he is leaving.

In the morning, he leaves home to go to his office. From his office, he goes for lunch at a restaurant, and then goes back to his office. After he is finished with his work at the office, he leaves for home. Suppose (for simplicity) that he does not go anywhere else. Suppose also that if he leaves it in the restaurant, it will remain there until the next time. Consider this situation as a Markov process and write down its matrix. What is the probability that after many days in the morning the umbrella is located in the restaurant? It is convenient to choose one day as a time unit: from morning to morning.

Solution.
$$A = \begin{pmatrix} 11/16 & 3/8 & 1/4 \\ 3/16 & 3/8 & 1/4 \\ 1/8 & 1/4 & 1/2 \end{pmatrix}$$

Compute the element a_1^1, that is, the probability that the umbrella starts its day at home and stays there, that is, it will be there the next morning. There are three distinct possibilities for the umbrella:

D the professor forgets it when leaving home in the morning $p_1 = \frac{1}{2}$,

DPD the professor takes it to the office, then he forgets to take it on to lunch and in the evening he takes it home: $p_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$,

DPRPD the professor takes the umbrella with him all the time and does not forget it anywhere: $p_3 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$.

In total $a_1^1 = p_1 + p_2 + p_3 = \frac{11}{16}$.

The eigenvector of this matrix corresponding to the dominant eigenvalue 1 is $(2, 1, 1)$, and thus the desired probability is $1/(2 + 1 + 1) = 1/4$. $\qquad\square$

3.D.5. 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 probabilities of transitions between pages are given by the hyperlink: each link, say from page

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_∞ of the matrix T with the eigenvalue 1, which is stochastic,*
- *the iterations $T^k x_0$ approach the vector x_∞ 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.

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_∞ 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, thus the only possibility is that $c_1 = 1$ and the second claim is proved.

A to page B, determines the probability (1/(total number of links from the page A)), with which the process moves from page A to page B. If from some page there are no leading links, we consider it to be a page from which a link leads to every other page. This gives a probabilistic matrix M (the element m_{ij} corresponds to the probability with which we move from the i-th page to the j-th page). Thus if one randomly clicks on links in the found pages (and from a linkless page one just chooses randomly the next one) the probability that at a given time (sufficiently large from the beginning) one is located on the i-th page corresponds to the i-th component of the unit eigenvector of the matrix M, corresponding to the eigenvalue 1. Looking at the sizes of these probabilities we define the importance of the individual pages.

This algorithm can be modified by assuming that users stop clicking from a link to a link after certain time and again starts on a random page. Suppose that with probability d he chooses a new page randomly, and with probability $(1 - d)$ keeps on clicking. In such a situation the probability of transition between any two pages S_i and S_j is non-zero – it is $d/n + (1 - d)$/total number of links at the page S_i if from S_i there is a link to S_j, and d/n otherwise (if there are no links at S_i, then it is $1/n$). According to the Perron-Frobenius theorem the eigenvalue 1 is with multiplicity one and dominant, and thus the corresponding eigenvector is unique (if we chose transitional probabilities only as described in the previous paragraph, it would not have to be so).

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, 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:

$$M = \begin{pmatrix} 1/20 & 1/20 & 17/20 & 1/4 \\ 9/20 & 1/20 & 1/20 & 1/4 \\ 9/20 & 17/20 & 1/20 & 1/4 \\ 1/20 & 1/20 & 1/20 & 1/4 \end{pmatrix}$$

The eigenvector corresponding to the eigenvalue 1 is $(305/53, 175/53, 315/53, 1)$, the importance of the pages is thus given according to the order of the sizes of the corresponding components, that is, $C > A > B > D$.

Another various applications of the Markov chains are in the additional exercises after this chapter, see

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 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. □

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 ε.

Corollary. *Let T be a primitive stochastic matrix from a Markov process and let x_∞ 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_∞, whose columns all equal to x_∞.*

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_∞. □

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_∞ 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:

$$T^2 = \begin{pmatrix} P^2 & P \cdot R + R \cdot Q \\ 0 & Q^2 \end{pmatrix}$$

$$T^3 = \begin{pmatrix} P^3 & P^2 \cdot R + P \cdot R \cdot Q + R \cdot Q^2 \\ 0 & Q^3 \end{pmatrix}.$$

E. Unitary spaces

In the previous chapter we defined the scalar product for real vector spaces (2.3.18). In this chapter we extend its definition to the complex spaces (3.4.1).

3.E.1. Groups $O(n)$ and $U(n)$. If we consider all linear mappings from $\mathbb{R}^3$ to $\mathbb{R}^3$ which preserve the given scalar product, that is, with respect to the definitions of the lengths of the vectors and deviations of two vector all linear mappings that preserve lengths and angles. Then these mappings form a group (see 1.1.1) with respect to the operation of composition. The composition of two such mappings is, by definition also a mapping that preserves lengths and angles, the unit element of the group is the identity mapping, and the inverse element for a given mapping is its inverse mapping. Such a mapping exists by the condition on the lengths preservation. The matrices of such mappings thus form a group with the operation of matrix multiplication (see); it is called the *orthogonal group* and is denoted by $0(n)$. It is a subgroup of the group of all invertible mappings from $\mathbb{R}^n$ to $\mathbb{R}^n$.

Moreover, if we require that the matrices have determinant one, then we speak of the special orthogonal group $SO(n)$. In general the determinant of a matrix in $O(n)$ can be either 1 or -1. Similarly we define the *unitary group $U(n)$* as the group of all (complex) matrices that correspond to the complex linear mappings from $\mathbb{C}^n$ to $\mathbb{C}^n$ which preserve a given scalar product in a unitary space. Analogously, $SU(n)$ denotes the subgroup of matrices in $U(n)$ with determinant one. In general, the determinant of a matrix in $U(n)$ can be any complex unit.

3.E.2. Consider the vector space V of functions $\mathbb{R} \to \mathbb{C}$. Determine whether the mapping φ from the unitary space V is linear when:

i) $\varphi(u) = \lambda u$ where $\lambda \in \mathbb{C}$
ii) $\varphi(u) = u^*$
iii) $\varphi(u) = u^2 (= u.u)$
iv) $\varphi(u) = \frac{du}{dx}$

For suitable functions V is a unitary space of infinite dimension. The scalar product is then defined by the relation $f.g = \int_{-\infty}^{\infty} f(x)\overline{g(x)}dx$.

3.E.3. 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 and compute its determinant.

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 $(E - Q)(E + Q + \cdots + Q^{k-1}) = E - Q^k$)

$$T^k = \begin{pmatrix} E & R(E - Q)^{-1}(E - Q^k) \\ 0 & Q^k \end{pmatrix}.$$

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 2.3.18–2.3.21) 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 α, 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:

<div style="text-align:center">UNITARY SPACES</div>

Unitary space is a complex vector space V along with the mapping $V \times V \to \mathbb{C}$, $(u, v) \mapsto u \cdot v$ 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.

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

$$U^* = (\sum_{n=0}^{\infty} \frac{1}{n!}(iH)^n)^* = \sum_{n=0}^{\infty} \frac{1}{n!}(-iH^*)^n$$

and since $H^* = H$, then

$$U^* = \sum_{n=0}^{\infty}(-1)^n \frac{1}{n!}(iH)^n = \exp(-iH).$$

Thus

$$U^*U = \exp(iH)\exp(-iH) = \exp(0) = 1.$$

$$\det(U) = \mathrm{e}^{\mathrm{trace}(iH)}.$$

$\square$

3.E.4. Hermitian matrices A, B, C satisfy $[A, C] = [B, C] = 0$ and $[A, B] \neq 0$, where $[,]$ is a commutator of matrices defined defined by the relation $[A, B] = AB - BA$. Show that at least one eigensubspace of the matrix C must have dimension > 1.

Solution. We prove it by contradiction. Assume that all eigensubspaces of the operator C have dim $= 1$. Then for any vector u we can write $u = \sum_k c_k u_k$ where u_k are linearly independent eigenvectors of the operator C associated with the eigenvalue λ_k (and $c_k = u.u_k$) For these eigenvectors

$$0 = [A, C]u_k = ACu_k - CAu_k = \lambda_k Au_k - C(Au_k).$$

From there it follows that Au_k is an eigenvector of the matrix C with the eigenvalue λ_k. But then that $Au_k = \lambda_k^A u_k$ for some number λ_k^A. Similarly, $Bu_k = \lambda_k^B u_k$ for some number λ_k^B. For the commutator of matrices A and B is then obtained

$$[A, B]u_k = ABu_k - BAu_k = \lambda_k^A \lambda_k^B u_k - \lambda_k^B \lambda_k^A u_k = 0$$

so that

$$[A, B]u = [A, B]\sum_k c_k u_k - \sum_k c_k[A, B]u_k = 0.$$

Because u is arbitrary, it follws that $[A, B] = 0$, which is a contradiction. $\square$

3.E.5. Applications to quantum physics. In quantum physics we do not use numbers as in classical physics, but a Hermitian operator. This is nothing but a Hermitian mapping, which can (and often does) lead to a linear transformation between unitary spaces of infinite dimension. We can imagine this as a matrix

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 $\mathbb{C}$:

$$u \cdot u \in \mathbb{R}$$
$$u \cdot u = 0 \quad \text{if and only if} \quad u = 0$$
$$u \cdot (av) = \bar{a}(u \cdot v)$$
$$u \cdot (v + w) = u \cdot v + u \cdot w$$
$$u \cdot 0 = 0 \cdot u = 0$$
$$(\sum_i a_i u_i) \cdot (\sum_j b_j v_j) = \sum_{i,j} a_i \bar{b}_j (u_i \cdot v_j),$$

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

$$(x_1, \ldots, x_n)^T \cdot (y_1, \ldots, x_n)^T = x_1 \bar{y}_1 + \cdots + x_n \bar{y}_n.$$

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 $\varphi : V \to W$ 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.

of infinite dimension. Vectors in this unitary space then represent the states of the given physical system. When measuring a given physical quantity we obtain only values that are eigenvalues of the corresponding operator.

For instance, instead of the coordinate x we have an operator of the coordinate $\hat{x}$, that results in multiplication by x. If the state of the system is described by the vector V, then $\hat{x}(v) = xv$. This corresponds to the multiplication of the vector by the real number x. At first glance this Hermitian operator is different from the cases of finite dimension. Evidently every real number is an eigenvalue and ($\hat{x}$ has a continuous spectrum). Similarly, instead of speed (more precisely, momentum) we have the operator $\hat{p} = -i\frac{d}{dx}$. The eigenvectors are the solution of the differential equation $-i\frac{dv}{dx} = \lambda v$. Even in this case the spectrum is continuous. This expresses the fact that the corresponding physical quantity is continuous and can attain any real value. On the other hand, we have physical quantities, for instance energy, that can attain only discrete values (energy exists in quanta). The corresponding operators are then really similar to the Hermitian matrices. They have infinitely many eigenvalues.

3.E.6. Show that $\hat{x}$ and $\hat{p}$ are Hermitian and that

$$[\hat{x}, \hat{p}] = i$$

Solution. For any vector v

$$[\hat{x}, \hat{p}]v = \hat{x}\hat{p}v - \hat{p}\hat{x}v = x(-i\frac{dv}{dx}) + i\frac{d(xv)}{dx} - = iv$$

from which the result follows. □

3.E.7. Show that

$$[\hat{x} - \hat{p}, \hat{x} + \hat{p}] = 2i$$

Solution. Evidently $[\hat{x}, \hat{x}] = 0$ and $[\hat{p}, \hat{p}] = 0$ The result follows from the linearity of the commutator from the previous exercise. □

3.E.8. Jordan form. Find the Jordan form of the following matrices. What is the geometric interpretation of this decomposition of the matrix?

i) $A = \begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix}$

ii) $A = \begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix}$

Athough 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 $(u_1, \ldots, u_k)$ there exists an orthonormal basis $(v_1, \ldots, v_n)$ such that $\langle v_1, \ldots, v_i \rangle = \langle u_1 \ldots, u_i \rangle$, for all $1 \le i \le k$, i.e. its vectors consecutively generate the same subspaces as the vector u_j.

(4) If $(u_1, \ldots, u_n)$ is an orthonormal basis V, then the coordinates of every vector $u \in V$ are expressed via

$$u = (u \cdot u_1)u_1 + \cdots + (u \cdot u_n)u_n.$$

(5) In any orthonormal basis, the scalar product has the coordinate form

$$u \cdot v = \bar{y}^T \cdot x = x_1\bar{y}_1 + \cdots + x_n\bar{y}_n$$

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 $V_1 + \cdots + V_k$ 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, $V = \operatorname{span} A \oplus A^\perp$.

(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 $(u_1, \ldots, u_n)$ of the space V and then start the Gramm-Schmidt orthogonalization from 2.3.20. This procedure works in the complex case. It yields an orthogonal basis with properties as required in (3). But from the Gramm-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 othogonalization process is applied. Thus we have also proved (2) and (1).

(4): If $u = a_1u_1 + \cdots + a_nu_n$, then

$$u \cdot u_i = a_1(u_1 \cdot u_i) + \cdots + a_n(u_n \cdot u_i) = a_i\|u_i\|^2 = a_i$$

(5): If $u = x_1u_1 + \cdots + x_nu_n, v = y_1u_1 + \cdots + y_nu_n$, then

$$u \cdot v = (x_1u_1 + \cdots + x_nu_n) \cdot (y_1u_1 + \cdots + y_nu_n)$$
$$= x_1\bar{y}_1 + \cdots + x_n\bar{y}_n.$$

Solution. i) First compute the characteristic polynomial of the matrix A

$$|A - \lambda E| = \begin{vmatrix} -1 - \lambda & 1 \\ -6 & 4 - \lambda \end{vmatrix} = \lambda^2 - 3\lambda + 2$$

The eigenvalues of the matrix A are the roots of this polynomial, that means that $\lambda_{1,2} = 1, 2$. Since the matrix is of order two, and has two distinct eigenvalues, its Jordan form is a diagonal matrix $J = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. The eigenvector (x, y) associated with the eigenvalue 1 satisfies $0 = (A - E)x = \begin{pmatrix} -2 & 1 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$, that is, $-2x + y = 0$. So the eigenvectors are the multiples of the vector $(1, 2)$.

Similarly the eigenvector associated with the eigenvalue 2 is $(1, 3)$. The matrix P is then obtained by writing these eigenvectors into tho columns, that is, $P = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$. For the matrix A, $A = P \cdot J \cdot P^{-1}$. The inverse of P is $P^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$, and

$$\begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$$

This decomposition says that the matrix A determines a linear mapping that has as basis of the eigenvectors $(1, 2)$, $(1, 3)$, the aforementioned diagonal form. Geometrically, this means that in the direction $(1, 2)$ nothing is changing and in the direction $(1, 3)$ every vector is being stretched twice.

ii) The characteristic polynomial of the matrix A is in this case

$$|A - \lambda E| = \begin{vmatrix} -1 - \lambda & 1 \\ -4 & 3 - \lambda \end{vmatrix} = \lambda^2 - 2\lambda + 1 = 0$$

There is a double root $\lambda = 1$ and the corresponding eigenvector (x, y) satisfies

$$0 = (A - E)x = \begin{pmatrix} -2 & 1 \\ -4 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

The solutions are, as in the previous case, multiples of the vector $(1, 2)$. The fact that the system does not have two linearly independent vectors as a solution says that the Jordan form in this case is not optimal, but it will be a matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. The basis for which A has this form is the eigenvector $(1, 2)$ and a vector that maps on this vector by the mapping $A - E$. Thus it is a solution of the system of equations

$$\begin{pmatrix} -2 & 1 & | & 1 \\ -4 & 2 & | & 2 \end{pmatrix} \sim \begin{pmatrix} -2 & 1 & | & 1 \\ 0 & 0 & | & 0 \end{pmatrix}$$

(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 $(v_1, \ldots, v_k)$ be a basis of span A chosen among the elements of A, and let $(u_1, \ldots, u_k)$ be the orthonormal basis resulting from the Gramm-Schmidt orthogonalization of the vectors $(v_1, \ldots, v_k)$. 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 span$\{u_{k+1}, \ldots, u_n\}$ = span$\{u_1, \ldots, u_k\}^\perp = A^\perp$ and $A \subset$ span$\{u_{k+1}, \ldots, u_n\}^\perp$ (this follows from expressing the coordinates under the orthonormal basis). If $u \perp$ span$\{u_{k+1}, \ldots, u_n\}$, then u is necessarily a linear combination of the vectors $u_1, \ldots, u_k$, but that happens whenever it is a linear combination of the vectors $v_1, \ldots, v_k$, which is equivalent to u being in span A.

(8): This is equivalent to the formulation of the existence of the orthonormal basis. $\qquad\square$

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.

<div align="center">PROPERTIES OF NORM</div>

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**.*

The solutions are multiples of the vector $(1, 3)$. We obtain the same basis as in the previous case and we can write

$$\begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$$

The mapping now acts on the vector as follows: the component in the direction $(1, 3)$ stays the same. The component in the direction $(1, 2)$ is multiplied by the sum of the coefficients that determine the components in the directions $(1, 3)$ and $(1, 2)$. □

3.E.9. Find the Jordan form of the matrices A_1 and A_2, and write down the decomposition. What is the geometric interpretation of this decomposition? $A_1 = \frac{1}{3} \begin{pmatrix} 5 & -1 \\ -2 & 4 \end{pmatrix}$ and $A_2 = \frac{1}{3} \begin{pmatrix} 5 & -1 \\ 4 & 1 \end{pmatrix}$ and show how the vectors $v = (3, 0)$, A_1v and A_2v decompose with respect to the basis of the eigenvectors of the matrix $A_{1,2}$.

Solution. The matrices have the same Jordan forms as the matrices in the previous exercise. In the basis of the vectors $(1, 2)$ and $(1, -1)$,

$$\frac{1}{3} \begin{pmatrix} 5 & -1 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}^{-1}$$

and

$$\frac{1}{3} \begin{pmatrix} 5 & -1 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}^{-1}$$

For the vector $v = (3, 0)$, $v = (1, 2) + 2(1, -1)$. For its images, $A_1v = (5, -2) = (1, 2) + 2 \cdot 2 \cdot (1, -1)$ and $A_2v = (5, 4) = (2 + 1) \cdot (1, 2) + 2 \cdot (1, -1)$. □

F. Matrix decompositions

3.F.1. Prove or disprove:

- Let A be a square matrix $n \times n$. Then the matrix $A^T A$ is symmetric.

- Let A be a square matrix with only real positive eigenvalues. Then A is symmetric.

3.F.2. Find an LU-decomposition of the following matrix:

$$\begin{pmatrix} -2 & 1 & 0 \\ -4 & 4 & 2 \\ -6 & 1 & -1 \end{pmatrix}$$

Solution.

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

First multiply the matrices that correspond to the Gaussian elimination, we thus obtain for the original matrix A, $XA =$

(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 $(e_1, \ldots, e_k)$ is a orthonormal system of vectors, then*

$$\|u\|^2 \geq |u \cdot e_1|^2 + \cdots + |u \cdot e_k|^2.$$

This property is called the **Bessel inequality**.

(4) *If $(e_1, \ldots, e_k)$ is an orthonormal system of vectors, then $u \in \operatorname{span}\{e_1, \ldots, e_k\}$ if and only if*

$$\|u\|^2 = |u \cdot e_1|^2 + \cdots + |u \cdot e_k|^2.$$

This is called the **Parseval equality**.

(5) *If $(e_1, \ldots, e_k)$ is an orthonormal system of vectors and $u \in V$, then the vector*

$$w = (u \cdot e_1)e_1 + \cdots + (u \cdot e_k)e_k$$

is the only vector which minimizes the norm $\|u - v\|$ among all $v \in \operatorname{span}\{e_1, \ldots, e_k\}$.

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

$$\|w\|^2 = \|u\|^2 - \frac{(\overline{u \cdot v})}{\|v\|^2}(u \cdot v) - \frac{u \cdot v}{\|v\|^2}(v \cdot u) + \frac{(u \cdot v)(\overline{u \cdot v})}{\|v\|^4}\|v\|^2$$

$$\|w\|^2\|v\|^2 = \|u\|^2\|v\|^2 - 2(u \cdot v)(\overline{u \cdot v}) + (u \cdot v)(\overline{u \cdot v})$$

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

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2 + u \cdot v + v \cdot u$$
$$= \|u\|^2 + \|v\|^2 + 2\operatorname{Re}(u \cdot v)$$
$$\leq \|u\|^2 + \|v\|^2 + 2|u \cdot v| \leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\|$$
$$= (\|u\| + \|v\|)^2$$

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 $(e_1, \ldots, e_k)$ be an orthonormal system of vectors. We extend it to an orthonormal basis $(e_1, \ldots, e_n)$ (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 $v \in \operatorname{span}\{e_1, \ldots, e_k\}$ and extend the given orthonormal system to the orthonormal basis $(e_1, \ldots, e_n)$. Let $(u_1, \ldots, u_n)$ and $(x_1, \ldots, x_k, 0, \ldots, 0)$ be

U, where X is a lower triangular matrix given by the Gaussian reduction, and U upper triangular. From this equality $A = X^{-1}U$, which is the desired decomposition. (Thus we have to compute the inverse of X). $\qquad\square$

3.F.3. Find the LU-decomposition of the matrix $\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \\ - & 1 & -1 \end{pmatrix}$. $\qquad\bigcirc$

3.F.4. 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 1.E.10.

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, $v_R = (-3, -2, -1)$. $\qquad\square$

3.F.5. Singular decomposition, polar decomposition, pseudoinverse. Compute the singular decomposition of the matrix

$$A = \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

. Then compute its polar decomposition and find its pseudoinverse.

Solution. First compute $A^T A$:

$$A^T A = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}$$

coordinates of u and v under this basis. Then

$$\|u-v\|^2 = |u_1-x_1|^2 + \cdots + |u_k-x_k|^2 + |u_{k+1}|^2 + \cdots + |u_n|^2$$

and this expression is clearly minimized when choosing the individual vectors to be $x_1 = u_1, \ldots, x_k = u_k$. $\qquad\square$

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) $\varphi : V \to V$ on the (real or complex) space with scalar product. Then the following conditions are equivalent.*

(1) φ is unitary or orthogonal transformation,
(2) φ 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 φ in any orthonormal basis satisfies $A^{-1} = \bar{A}^T$ (for Euclidean spaces this means that $A^{-1} = A^T$),
(4) the matrix A of a mapping φ in some orthonormal basis satisfies $A^{-1} = \bar{A}^T$,
(5) the rows of the matrix A of the mapping φ 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 φ in an orthonormal basis form an orthonormal basis of the space $\mathbb{K}^n$ with standard scalar product.

PROOF. $(1) \Rightarrow (2)$: The mapping φ 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 φ is invertible and $\bar{y}^T A x = (\overline{A^{-1}y})^T x$. This means that $(\bar{y}^T A - (\overline{A^{-1}y})^T)x = 0$ 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 φ as the equation $A\bar{A}^T = E$, which is ensured by (4).
$(5) \Rightarrow (6)$: We have $|\bar{A}^T A| = |E| = |A\bar{A}^T| = |A||\overline{|A|} = 1$, 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

$$\varphi(u) \cdot \varphi(v) = \overline{(Ay)}^T Ax = \bar{y}^T \bar{A}^T Ax = \bar{y}^T \bar{E}x = \bar{y}^T x$$

where x and y are columns of coordinates of the vectors u and v. That ensures that the scalar product is preserved. $\qquad\square$

to obtain a diagonal matrix. We need to find an orthonormal basis under which the matrix is diagonal and the zero row is the last one. This can be obtained by rotating about the x-axis through a right angle. The y-coordinate then goes to z and z goes to $-y$. This rotation is an orthogonal transformation given by the matrix $V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$. By this, we have found the decomposition $A^T A = V B V^T$. here, B is diagonal with eigenvalues $(1, \frac{1}{4}, 0)$ on the diagonal. Because $B = (AV)^T(AV)$, the columns of the matrix

$$AV = \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

form an orthogonal system of vectors, which we normalise and extend to a basis. That is then of the form $(0, -1, 0), (1, 0, 0), (0, 0, 1)$. The transition matrix of changing from this basis to the standard one is then

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

Finally, we obtain the decomposition $A = U\sqrt{B}V^T$

$$\begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

The geometrical interpretation of decomposition is the following: first, everything is rotated through a right angle by the x-axis, then follows a projection to the xy plane such that the unit ball is mapped on the ellipse with major half-axes 1 and $\frac{1}{2}$. The result is then rotated through a right angle about the z-axis.

The polar decomposition $A = P \cdot W$ can be obtained from the singular one: $P := U\sqrt{B}U^T$ and $W := UV^T$, that is,

$$P = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

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

Characterizations from the previous theorem deserve some notes. The matrices $A \in \mathrm{Mat}_n(\mathbb{K})$ 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 $U(n) \subset \mathrm{Gl}_n(\mathbb{C})$ in the group of all invertible complex matrices with the product operation. Orthogonal matrices form a subgroup $O(n) \subset \mathrm{Gl}_n(\mathbb{R})$ 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 ± 1. Furthermore, if $Ax = \lambda x$ for a unitary or orthogonal matrix, then $(Ax) \cdot (Ax) = x \cdot x = |\lambda|^2(x \cdot x)$. Therefore the real eigenvalues of orthogonal matrices in the real domain are ± 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 $\varphi : V \to V$ 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 $\varphi : V \to V$ 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 φ 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. □

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.

From this it follows that

$$\begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

The pseudoinverse matrix is then given by the expression $A^{(-1)} := VS'U^T$, where $S' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Thus,

$$A^{(-1)} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

$\square$

3.F.6. QR decomposition. The QR decomposition of a matrix A is very useful when we are given a system of linear equations $Ax = b$ which has no solution, but an approximation as good as possible is needed. That is, we want to minimize $\|Ax - b\|$. According to the Pythagorean theorem, $\|Ax - b\|^2 = \|Ax - b_\|\|^2 + \|b_\perp\|^2$, where b is decomposed into $b_\|$ which belongs to the range of the linear transformation A, and into $b_\perp$, which is perpendicular to this range. The projection on the range of A can be written in the form QQ^T for a suitable matrix Q. Specifically for this matrix we obtain it by the Gram-Schmidt orthonormalisation of the columns of the matrix A. Then $Ax - b_\| = Q(Q^TAx - Q^Tb)$. The system in the parentheses has a solution, for which $\|Ax - b\| = \|b_\perp\|$, which is the minimal value. Furthermore, the matrix $R := Q^TA$ is upper triangular and therefore the approximate solution can be found easily.

Find an approximate solution of the system

$$x + 2y = 1$$
$$2x + 4y = 4$$

Solution. Consider the system $Ax = b$ with $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$, which evidently has no solution. We orthonormalise the columns of A. We take the first of them and divide it by its norm. This yields the first vector of the orthonormal basis $\frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. But the second is twice the first and thus it will be after orthonormalisation. Therefore $Q = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

The projector on the range of A is then $QQ^T = \frac{1}{5}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$.

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 $\psi : V \to W$, the expression

$$(1) \qquad \langle v, \psi^*(\alpha) \rangle = \langle \psi(v), \alpha \rangle,$$

where $\langle\ ,\ \rangle$ 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 $\psi^* : W^* \to V^*$ called the *dual mapping* to ψ.

Choose bases $\underline{v}$ in V, $\underline{w}$ in W and write A for the matrix of the mapping ψ in these bases. Then we compute the matrix of the mapping ψ^* in the corresponding dual bases in the dual spaces. Indeed, the definition says that if we represent the vectors from W^* in the coordinates as rows of scalars, then the mapping ψ^* is given by the same matrix as ψ, 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 ψ^* 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 $W^* \to V^*$ look in terms of our identification? We use the same notation $\psi^* : W \to V$ for the resulting mapping, which is uniquely given as follows:

Next,
$$Q^T b = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{9}{\sqrt{5}}$$
and
$$R = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 5 & 9 \end{pmatrix}.$$
The approximate solution then satisfies $Rx = Q^T b$, and here that means $5x + 9y = 9$. (The approximate solution is not unique). The QR decomposition of the matrix A is then
$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \frac{1}{\sqrt{5}} \begin{pmatrix} 5 & 9 \end{pmatrix}$$
□

3.F.7. Minimise $\|Ax - b\|$ for $A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Hence write down the QR decomposition of the matrix A.

Solution. The normalised first column of the matrix A is $e_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$. From the second column, subtract its component in the direction e_1. Then
$$\left\langle \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \right\rangle = -\frac{3}{\sqrt{6}}$$
and therefore
$$\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} - \left\langle \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \right\rangle \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$$
By this we have created an orthogonal vector, which we normalise to obtain $e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$. The third column of the matrix A is already linearly dependent (verify this by computing the determinant, or otherwise). The desired column-orthogonal matrix is then
$$Q = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & 0 \\ -1 & \sqrt{3} \\ -1 & -\sqrt{3} \end{pmatrix}$$
Next,
$$R = Q^T A = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$
$$= \frac{1}{\sqrt{6}} \begin{pmatrix} 6 & -3 & -3 \\ 0 & 3\sqrt{3} & -3\sqrt{3} \end{pmatrix}$$
and
$$Q^T b = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

ADJOINT MAPPING

For every linear mapping $\psi : V \to W$ between spaces with scalar products, there is the *adjoint mapping* ψ^* uniquely determined by the formula

(2) $\qquad \langle \psi(u), v \rangle = \langle u, \psi^*(v) \rangle.$

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 ψ^*. 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 ψ^*.

Using the coordinate expression for the scalar product, the formula (2) reveals the coordinate expression of the adjoint mapping:
$$\langle \psi(v), w \rangle = \overline{(w_1, \ldots, w_n)} \cdot A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$
$$= \overline{\left(\bar{A}^T \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \right)}^T \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \langle v, \psi^*(w) \rangle.$$

It follows that if A is the matrix of the mapping ψ in an orthonormal basis, then the matrix of the adjoint mapping ψ^* 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 $\psi : V \to W$ 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 $\psi^* : W \to V$, 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.

The solution of the equation $Rx = Q^T b$ is $x = y = z$. Thus, multiples of the vector $(1, 1, 1)$ minimize $\|Ax - b\|$.

The mapping given by the matrix A is a projection on the plane with normal vector $(1, 1, 1)$.

$\square$

3.F.8. Linear regression. The knowledge obtained in this chapter can be successfully used in practice for solving problems with linear regression. It is about finding the best approximation of some functional dependence using a linear function.

Given a functional dependence for some points that is, $f(a_1^1, \ldots, a_n^1) = y_1, \ldots, f(a_1^k, a_2^k, \ldots, a_n^k) = y_k$, $k > n$ (we have thus more equations than unknowns) and we wish to find the "best possible" approximation of this dependency using a linear function. That is, we want to express the value of the property as a linear function $f(x_1, \ldots, x_n) = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + c$. 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 $b_1, \ldots, b_n, c$. The goal is to find such a linear combination of the columns of the matrix $A = (a_j^i)$ (with coefficients $b_1, \ldots, b_n$), that is closest to the vector $(y_1, \ldots, y_k)$ in $\mathbb{R}^k$. Thus it is about finding an orthogonal projection of the vector $(y_1, \ldots, y_k)$ on the subspace generated by the columns of the matrix A. Using the theorem 3.5.7 this projection is the vector $(b_1, \ldots, b_n)^T = A^{(-1)}(y_1, \ldots, b_n)$.

3.F.9. Using the least squares method, solve the system

$$
\begin{aligned}
2x + y + 2z &= 1 \\
x + y + 3z &= 2 \\
2x + y + z &= 0 \\
x + z &= -1
\end{aligned}
$$

Solution. The system has no solution, since its matrix has rank 3, and the extended matrix has rank 4. The best approximation of the vector $b = (1, 2, 0, -1)$ can thus be obtained using the theorem 3.5.7 by the vector $A^{(-1)}b$. $AA^{(-1)}b$ is then the best approximation – the perpendicular projection

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 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

$$A = B + i\,C$$

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

$$B = \operatorname{re} A = \frac{1}{2}(A + A^*), \; C = \operatorname{im} A = \frac{1}{2i}(A - A^*).$$

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 $\psi : V \to V$ 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 $\psi : V \to V$ 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

$$\lambda \langle x, x \rangle = \langle Ax, x \rangle = \langle x, Ax \rangle = \langle x, \lambda x \rangle = \bar{\lambda} \langle x, x \rangle.$$

of the vector b on the space generated by the columns of the matrix A.

Because the columns of the matrix A are linearly independent, its pseudoinverse is given by the relation $(A^T A)^{-1} A^T$. Hence

$A^{(-1)}$

$$= \left(\begin{pmatrix} 2 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \left(\begin{pmatrix} 10 & 5 & 10 \\ 5 & 3 & 6 \\ 10 & 6 & 15 \end{pmatrix} \right)^{-1} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 3/5 & -1 & 0 \\ -1 & 10/3 & -2/3 \\ 0 & -2/3 & 1/3 \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1/5 & -2/5 & 1/5 & 3/5 \\ 0 & 1/3 & 2/3 & -5/3 \\ 0 & 1/3 & -1/3 & 1/3 \end{pmatrix}$$

The desired $\mathbf{x}$ is

$$A^{(-1)} b = (-6/5, 7/3, 1/3)^T.$$

The projection (the best possible approximation to the column of the right side) is then the vector $(3/5, 32/15, 4/15, -13/15)$. $\qquad \square$

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 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 $V_1 \subset V_2 \subset \ldots V_n = V$, 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 $P : V \to V$ 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 $P : V \to V$ 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 $\psi : V \to V$ on a vector space with scalar product there exists an orthonormal basis composed of eigenvectors. If $\lambda_1, \ldots, \lambda_k$ are all distinct eigenvalues of ψ and if $P_1, \ldots, P_k$ are the corresponding orthogonal and mutually perpendicular projectors onto the eigenspaces corresponding to the eigenvalues, then*

$$\psi = \lambda_1 P_1 + \cdots + \lambda_k P_k.$$

The dimensions of the images of these projections P_i equal the algebraic multiplicities of the eigenvalues λ_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 ± 1.

The situation is much more interesting on unitary spaces. Consider any linear mapping $\varphi : V \to V$ on a unitary space. Let $\varphi = \psi + i\eta$ be the (unique) decomposition of φ into its Hermitian and anti-Hermitian part. If φ has diagonal matrix D in a suitable orthonormal basis, then $D = \operatorname{Re} D + i \operatorname{Im} D$, where the real and the imaginary parts are exactly the matrices of ψ and η. This follows from the uniqueness of the decomposition. Knowing this in the particular coordinates, we conclude the following computation relations at the level of mappings $\psi \circ \eta = \eta \circ \psi$ (i.e. the real and imaginary parts of φ commute), and $\varphi \circ \varphi^* = \varphi^* \circ \varphi$ (since this clearly holds for all diagonal metrices). The mappings $\varphi : V \to V$ 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 $\varphi : V \to V$ on a unitary space V are equivalent:*

(1) φ is orthogonally diagonalizable,

(2) $\varphi^* \circ \varphi = \varphi \circ \varphi^*$ (φ 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 φ in some orthonormal basis, and λ_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 φ is normal, then

$$\langle \varphi(u), \varphi(u) \rangle = \langle \varphi^*\varphi(u), u \rangle = \langle \varphi\varphi^*(u), u \rangle$$
$$= \langle \varphi^*(u), \varphi^*(u) \rangle$$

thus $|\varphi(u)| = |\varphi^*(u)|$.

Next, notice $(\varphi - \lambda \operatorname{id} V)^* = (\varphi^* - \bar{\lambda} \operatorname{id} V)$. Thus, if φ is normal, then $(\varphi - \lambda \operatorname{id} V)$ is normal too.

If $\varphi(u) = \lambda u$, then u is in the kernel of $\varphi - \lambda \operatorname{id}_V$. Thus the latter equality of norms of values for normal mappings and their adjoints ensures that u is also in the kernel of $\varphi^* - \bar{\lambda} \operatorname{id}_V$. It follows that $\varphi^*(u) = \bar{\lambda} u$. We have proved, under the assumption (2), that φ and φ^* 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 λ, 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 φ 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 linear mapping $V \to V$, which we prove later in 3.4.15. This theorem says that for every linear mapping $\varphi : V \to V$ there exists an orthonormal basis under which φ has an upper triangular matrix. Then all the eigenvalues of φ 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. $\square$

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 φ on V for which the unique decomposition $\varphi = \psi + i\eta$ into Hermitian and anti-Hermitian parts satisfies $\psi^2 + \eta^2 = \operatorname{id}_V$.*

PROOF. If φ is unitary, then $\varphi\varphi^* = \operatorname{id}_V = \varphi^*\varphi$ and thus $\varphi\varphi^* = (\psi + i\eta)(\psi - i\eta) = \psi^2 + 0 + \eta^2 = \operatorname{id}_V$. On the other hand, if φ is normal, we can read the latter computation backwards which proves the other implication. $\square$

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 φ with their adjoints.

By definition,

$$(1) \qquad \langle B\,x, x \rangle = \langle A^*A\,x, x \rangle = \langle A\,x, A\,x \rangle \geq 0$$

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 $\varphi : V \to V$.

For every mapping $\varphi : V \to V$ we can define its *square root* as a mapping ψ such that $\psi \circ \psi = \varphi$. The next theorem completely describes the situation when restricting to positive semidefinite mappings.

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 φ, 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 ψ given by C must have the same eigenvectors as φ 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 φ. □

Notice there could be a lot of different roots, if we relax the positivity condition on $\sqrt{B}$, see **??**.

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* $f : V \to V$ 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 $f : V \to V$ 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 $f : V \to V$ 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

$$
A = \begin{pmatrix} 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & & \ddots \end{pmatrix}.
$$

If $f(v) = a\,v$, 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 λ given by the matrix

$$
(1) \qquad J = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & \lambda \end{pmatrix}.
$$

These matrices (and the corresponding invariant subspaces) are called *Jordan blocks*.[4]

---

[4]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.

Theorem. *Let V be a real or complex vector space of dimension n. Let $f : V \to V$ 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*

$$V = V_1 \oplus \cdots \oplus V_k$$

where not only $f(V_i) \subset V_i$, but the restriction of f to each V_i has a single eigenvalue λ_i and the restriction $f - \lambda_i \operatorname{id}_{V_i}$ 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 $A = P^{-1} J P$ 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 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 eaxactly 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 λ on all the complex space V and that $f - \lambda \operatorname{id}_V$ is cyclic (that is, we consider only one Jordan block separately). Let $v_1, \ldots, v_\ell$ 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 λ 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 $\varphi : V \to V$, if there exists an $a \in \mathbb{K}$ and an integer $k > 0$ such that $(\varphi - a\,\mathrm{id}_V)^k(u) = 0$. This means that the k-th iteration of the given mapping sends u to zero. The set of all root vectors corresponding to a fixed scalar λ 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 $(\varphi - a\,\mathrm{id}_V)^{k-1}(u)$ is an eigenvector with the eigenvalue a. Thus we have $\mathcal{R}_\lambda = \{0\}$ for all scalars λ which are not in the spectrum of the mapping φ.

Proposition. *Let $\varphi : V \to V$ 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 $(\varphi - \mu\,\mathrm{id}_V)$. In particular $\mathcal{R}_\lambda$ is invariant with respect to φ,

(3) if $\mu \neq \lambda$, then $(\varphi - \mu\,\mathrm{id}_V)|_{\mathcal{R}_\lambda}$ is invertible,

(4) the mapping $(\varphi - \lambda\,\mathrm{id}_V)|_{\mathcal{R}_\lambda}$ is nilpotent.

PROOF. (1) Checking the properties of the vector vector subspace is easy and is left to the reader.
(2) Assume that $(\varphi - \lambda\,\mathrm{id}_V)^k(u) = 0$ and put $v = (\varphi - \mu\,\mathrm{id}_V)(u)$. Then

$$(\varphi - \lambda\,\mathrm{id}_V)^k(v) =$$
$$= (\varphi - \lambda\,\mathrm{id}_V)^k((\varphi - \lambda\,\mathrm{id}_V) + (\lambda - \mu)\,\mathrm{id}_V)(u)$$
$$= (\varphi - \lambda\,\mathrm{id}_V)^{k+1}(u) + (\lambda - \mu) \cdot (\varphi - \lambda\,\mathrm{id}_V)^k(u) = 0$$

(3) If $u \in \mathrm{Ker}(\varphi - \mu\,\mathrm{id}_V)|_{\mathcal{R}_\lambda}$, then

$$(\varphi - \lambda\,\mathrm{id}_V)(u) = (\varphi - \mu\,\mathrm{id}_V)(u) + (\mu - \lambda)\,u = (\mu - \lambda)\,u.$$

This implies $0 = (\varphi - \lambda \, \mathrm{id}_V)^k(u) = (\mu - \lambda)^k \, u$ and thus also $u = 0$ for $\lambda \neq \mu$.

(4) Choose a basis $e_1, \ldots, e_p$ of the subspace $\mathcal{R}_\lambda$. By definition, there exist integers k_i such that $(\varphi - \lambda \, \mathrm{id}_V)^{k_i}(e_i) = 0$. In particular, the entire mapping $(\varphi - \lambda \, \mathrm{id}_V)|_{\mathcal{R}_\lambda}$ must be nilpotent. $\qquad\square$

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.

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 $[v] + [w] = [v+w]$, $a\,[u] = [a\,u]$. 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 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 $(u_1, \ldots, u_n)$ be a basis of V, such that $(u_1, \ldots, u_k)$ is a basis of U. Then $\dim V/U = n - k$ and the vectors*

$$u_{k+1} + U, \ldots, u_n + U$$

form a basis of V/U.

PROOF. $V = \operatorname{span}\{u_1, \ldots, u_n\}$, so $V/U = \operatorname{span}\{u_1 + U, \ldots, u_n + U\}$. But the first k generators are zero, thus $V/U = \operatorname{span}\{u_{k+1} + U, \ldots, u_n + U\}$. Assume that the linear combination $a_{k+1}(u_{k+1} + U) + \cdots + a_n(u_n + U) = (a_{k+1} u_{k+1} + \cdots + a_n u_n) + U = 0 \in V/U$ vanishes. Equivalently, this linear combination of the vectors $u_{k+1}, \ldots, u_n$ 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. $\quad\square$

3.4.14. Induced mappings on quotient spaces. Assume that $U \subset V$ is an invariant subspace with respect to linear mapping $\varphi : V \to V$ and choose basis $u_1, \ldots, u_n$ of the space V such that the first k vectors of this basis is a basis of U. With this basis, φ has block matrix $A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$. Then we can prove the following lemma:

Lemma. *(1) the mapping φ induces a linear mapping $\varphi_{V/U} : V/U \to V/U$, $\varphi_{V/U}(v+U) = \varphi(v)+U$ with the matrix D under the induced basis $u_{k+1}+U, \ldots, u_n+U$ on V/U,*

(2) the characteristic polynomial of $\varphi_{V/U}$ divides the characteristic polynomial of φ.

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 $a(\varphi(v)+U) = a\varphi(v)+U = \varphi(a\,v)+U$ (because φ 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 φ is $|A - \lambda\, E| = |B - \lambda\, E||D - \lambda\, E|$. $\square$

Corollary. *Let V be a vector space over $\mathbb{K}$ of dimension n and let $\varphi : V \to V$ 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 $\{0\} = V_0 \subset V_1 \subset \cdots \subset V_n = V$ with dimensions $\dim V_i = i$. Consider a basis $u_1, \ldots, u_n$ of the space V such that $V_i = \mathrm{span}\{u_1, \ldots, u_i\}$. In this basis, the matrix of the mapping φ is an upper triangular matrix:*

$$\begin{pmatrix} \lambda_1 & \ldots & * \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix},$$

with the spectrum $\lambda_1, \ldots, \lambda_n$ on the diagonal.

PROOF. The subspaces V_i are constructed inductively. Let $\{\lambda_1, \ldots, \lambda_n\}$ be the spectrum of the mapping φ. Thus the characteristic polynomial of the mapping φ is of the form $\pm(\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$. We choose $V_0 = \{0\}$, $V_1 = \mathrm{span}\{u_1\}$, where u_1 is an eigenvector with eigenvalue λ_1. According to the previous theorem, the characteristic polynomial of the mapping φ_{V/V_1} is of the form $\pm(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$. Assume that we have already constructed linearly independent vectors $u_1, \ldots, u_k$ and invariant subspaces $V_i = \mathrm{span}\{u_1 \ldots, u_i\}$, $i = 1, \ldots, k < n$ such that the characteristic polynomial of φ_{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 $i = 1, \ldots, k$.

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 φ_{V/V_k} with the eigenvalue λ_{k+1}. Consider the space $V_{k+1} = \operatorname{span}\{u_1, \ldots, u_{k+1}\}$. If the vector u_{k+1} is a linear combination of the vectors $u_1, \ldots, u_k$ 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 φ. But completing the vectors $u_1, \ldots, u_{k+1}$ to the basis of V yields a block matrix of the mapping φ 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 $\lambda_1, \ldots, \lambda_{k+1}$. Therefore the roots of the characteristic polynomial of the induced mapping have the required properties. $\qquad\square$

Remark. If V decomposes into the direct sum of eigensubspaces for φ, 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 φ 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 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 φ_{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:

<small>SCHUR'S ORTHOGONAL TRIANGULATION THEOREM</small>

Theorem. *Let $\varphi : V \to V$ 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 φ in this basis is upper triangular with eigenvalues $\lambda_1, \ldots, \lambda_m$ on the diagonal.*

3.4.16. Theorem. *Let $\varphi : V \to V$ be a linear mapping and $\lambda_1, \ldots, \lambda_k$ be all distinct eigenvalues. Then the sum of the*

root spaces $\mathcal{R}_{\lambda_1}, \ldots, \mathcal{R}_{\lambda_k}$ is direct. Furthermore, for every eigenvalue λ the dimension of the subspace $\mathcal{R}_\lambda$ equals the algebraic multiplicity of λ.

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 $u_1 \in \mathcal{R}_{\lambda_1}, \ldots, u_k \in \mathcal{R}_{\lambda_k}$ satisfy $u_1 + \cdots + u_k = 0$. Then, $(\varphi - \lambda_k \, \mathrm{id}_V)^j(u_k) = 0$ for suitable j, and moreover all $y_i = (\varphi - \lambda_k \, \mathrm{id}_V)^j(u_i)$ are non-zero vectors in $\mathcal{R}_{\lambda_i}$, $i = 1, \ldots, k - 1$, whenever u_i are non-zero by Proposition 3.4.12. But at the same time

$$ y_1 + \cdots + y_{k-1} = (\varphi - \lambda_k \cdot \mathrm{id}_V)^j \left(\sum_{i=1}^k u_i \right) = 0 $$

and, according to the inductive assumption, all y_i are zero. But then also all u_i, $1 \le i < k$ 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 λ of φ, use the same notation φ for the restriction $\varphi|_{\mathcal{R}_\lambda}$ and write $\psi : V/\mathcal{R}_\lambda \to V/\mathcal{R}_\lambda$ for the mapping induced by φ on the quotient space.

Assume that the dimension $\mathcal{R}_\lambda$ is strictly smaller than the algebraic multiplicity of the root λ of the characteristic polynomial. In view of lemma 3.4.14, λ is also an eigenvalue of the mapping ψ. 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 $w = (\varphi - \lambda \, \mathrm{id}_V)(v)$ and $(\varphi - \lambda \, \mathrm{id}_V)^j(w) = 0$ for suitable j. We conclude that $(\varphi - \lambda \, \mathrm{id}_V)^{j+1}(v) = 0$, 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 λ of the characteristic polynomial of the mapping $\varphi : V \to V$. □

Combining the latter theorem with the triangulation result from Corollary 3.4.14, we can formulate:

Corollary. *Consider a linear mapping $\varphi : V \to V$ on a vector space V over scalars $\mathbb{K}$, whose entire spectrum is in $\mathbb{K}$. Then $V = \mathcal{R}_{\lambda_1} \oplus \cdots \oplus \mathcal{R}_{\lambda_n}$ is the direct sum of the root subspaces. If we choose suitable bases for these subspaces, then under this basis φ has block-diagonal form with upper triangular matrices in the blocks and eigenvalues λ_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 $\varphi : V \to V$ be a nilpotent linear mapping. Then there exists a decomposition of V into a direct sum of subspaces $V = V_1 \oplus \cdots \oplus V_k$ such that the restriction of φ to each summand V_i is cyclic.*

188

PROOF. We provide a straightforward construction of a basis of the space V such that the action of the mapping φ on the basis vectors directly shows the decomposition into the cyclic mappings.

Let k be the degree of nilpotency of the mapping φ and write $P_i = \operatorname{Im}(\varphi^i)$, $i = 0, \ldots, k$. Thus,

$$\{0\} = P_k \subset P_{k-1} \subset \cdots \subset P_1 \subset P_0 = V.$$

Choose a basis $e_1^{k-1}, \ldots, e_{p_{k-1}}^{k-1}$ of the space P_{k-1}, where $p_{k-1} > 0$ is the dimension of P_{k-1}. By definition, $P_{k-1} \subset \operatorname{Ker}\varphi$, 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}, $j = 1, \ldots, p_{k-1}$ in P_{k-2}, such that $\varphi(e_j^{k-2}) = e_j^{k-1}$. Assume

$$a_1 e_1^{k-1} + \cdots + a_{p_{k-1}} e_{p_{k-1}}^{k-1} + b_1 e_1^{k-2} + \cdots + b_{p_{k-1}} e_{p_{k-1}}^{k-2} = 0.$$

Applying φ on this linear combination yields $b_1 e_1^{k-1} + \cdots + b_{p_{k-1}} e_{p_{k-1}}^{k-1} = 0$. 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)
$$\begin{aligned} &e_1^{k-1}, \ldots, e_{p_{k-1}}^{k-1} \\ &e_1^{k-2}, \ldots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \ldots, e_{p_{k-2}}^{k-2} \end{aligned}$$

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 the basis elements $e_1^{k-1}, \ldots, e_{p_{k-1}}^{k-1}$. We can thus adjust the chosen vectors $e_{p_{k-1}+1}^{k-2}, \ldots, e_{p_{k-2}}^{k-2}$ by adding the appropriate linear combinations of the vectors $e_1^{k-2}, \ldots, e_{p_{k-1}}^{k-2}$ with the result that they are in the kernel of φ. 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)

$$\begin{aligned} &e_1^{k-1}, \ldots, e_{p_{k-1}}^{k-1} \\ &e_1^{k-2}, \ldots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \ldots, e_{p_{k-2}}^{k-2} \\ &e_1^{k-3}, \ldots, e_{p_{k-1}}^{k-3}, e_{p_{k-1}+1}^{k-3}, \ldots, e_{p_{k-2}}^{k-3}, e_{p_{k-2}+1}^{k-3}, \ldots, e_{p_{k-3}}^{k-3} \\ &\vdots \\ &e_1^{k-\ell}, \ldots, e_{p_{k-1}}^{k-\ell}, e_{p_{k-1}+1}^{k-\ell}, \ldots, e_{p_{k-2}}^{k-\ell}, e_{p_{k-2}+1}^{k-\ell}, \ldots\ldots e_{p_{k-\ell}}^{k-\ell} \end{aligned}$$

where the value of the mapping φ 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 $e_1^{k-\ell-1}, \ldots, e_{p_{k-\ell}}^{k-\ell-1}$ which map to $e_1^{k-\ell}, \ldots, e_{p_{k-\ell}}^{k-\ell}$. We can extend them to a basis P_{k-l-1}, say, by the vectors

$$e_{p_{k-\ell}+1}^{k-\ell-1}, \ldots, e_{p_{k-\ell-1}}^{k-\ell-1}.$$

Again, exactly as when adjusting (1) above, we choose the additional basis vectors from the kernel of φ. and analogically as before we verify that we indeed obtain a basis for $P_{k-\ell-1}$.

189

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 φ are cyclic mappings. $\qquad\square$

3.4.18. Proof of the Jordan theorem. Let $\lambda_1, \dots, \lambda_k$ be all the distinct eigenvalues of the mapping φ. From the assumptions of the Jordan theorem it follows that $V = \mathcal{R}_{\lambda_1} \oplus \cdots \oplus \mathcal{R}_{\lambda_k}$.

The mappings $\varphi_i = (\varphi|_{\mathcal{R}_{\lambda_i}} - \lambda_i \,\mathrm{id}_{\mathcal{R}_{\lambda_i}})$ are nilpotent and thus each of the root spaces is a direct sum

$$\mathcal{R}_{\lambda_i} = P_{1,\lambda_i} \oplus \cdots \oplus P_{j_i,\lambda_i}$$

of spaces on which the restriction of the mapping $\varphi - \lambda_i \,\mathrm{id}_V$ 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 λ_i.

For the proof of Jordan theorem it remains to verify the claim about uniqueness (up to reordering the blocks). Because the diagonal values λ_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 λ and we are going to express the dimensions of individual Jordan blocks using the ranks r_k of the mapping $(\varphi - \lambda \,\mathrm{id}_V)^k$. 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 ψ is a cyclic operator on an m-dimensional space, then the defect of the iterated mapping ψ^k is k for $0 \le k \le m$, while the defect is m for all $k \ge m$. This implies that if our matrix J of the mapping φ 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_ℓ'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 $\varphi : V \to V$ has its matrix in the canonical Jordan form.[5]

(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 φ has diagonal matrix.

(4) Let λ be the eigenvalue with geometric multiplicity strictly smaller than the algebraic multiplicity and $v_1, \ldots, v_k$ 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 $\varphi - \lambda\,\mathrm{id}_V$. 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 $(\varphi - \lambda\,\mathrm{id})(w_i) = v_i$ 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 $\varphi - \lambda\,\mathrm{id}$ is cyclic (the columns from the scheme in the proof).

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

$$\mathrm{span}\{(1,0,0)^T, (0,1,0)^T\},$$

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]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.

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

$$a = \operatorname{sgn}(a) \cdot |a|,$$

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 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, $A = L\,D\,U$.

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!)

<div align="center">

LU-DECOMPOSITION

</div>

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 $f : V \to W$ 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 $f : V \to V$ over complex vector spaces. We 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 $f : V \to W$ 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 $A A^*$.

If A is a real matrix, then the matrices U and V are orthogonal.

PROOF. Assume first that $m \leq n$. Denote by $\varphi : \mathbb{K}^n \to \mathbb{K}^m$ 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 φ 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 $\underline{w}$ 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 $A^*A = V B V^*$ for a real diagonal matrix B with non-negative eigenvalues $(d_1, d_2, \ldots, d_r, 0, \ldots, 0)$ on the diagonal, $d_i \neq 0$ for all $i = 1, \ldots, r$. Thus

$$B = V^* A^* A V = (A V)^*(A V).$$

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 $v_1, \ldots, v_r \in \mathbb{K}^m$. Thus, $\langle v_i, v_i \rangle = d_i$, $i = 1, \ldots, r$, 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 $\underline{u} = u_1, \ldots, u_m$ for the entire $\mathbb{K}^m$. Expressing the original mapping φ in the bases $\underline{w}$ 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. $\qquad\square$

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 $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ 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 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:

<center>POLAR DECOMPOSITION</center>

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.

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:

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 $\varphi : \mathbb{K}^n \to \mathbb{K}^m$ with the matrix A in the standard bases we can choose a new orthonormal basis on $\mathbb{K}^m$ for which φ 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 Gramm-Schmidt orthogonalization of this system of vectors.

For every e_i, $\varphi(e_i)$ is either one of v_j, $j \le i$, 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 φ 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.

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.

Let A be a real or complex matrix of the type m/n. Let

$$A = U S V^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

be its singular decomposition (in particular, D is invertible). The matrix

$$A^\dagger := V S^\dagger U^*, \quad S^\dagger = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

is called the *pseudoinverse matrix* of the matrix A.

In geometric terms, we may view the linear mapping φ given by the matrix A in the two special orthonormal basis, where φ has got the matrix S with non-negative diagonal entries. We take the inverse of the "invertible part" of φ and

complete it trivially to the pseudoinverse[6] 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.

<div align="center">PROPERTIES OF PSEUDOINVERSE MATRICES</div>

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

$$AA^\dagger A = USV^*VS^\dagger U^*USV^* = USS^\dagger SV^* = USV^* = A$$

and analogically for the second equation. Furthermore,

$$(AA^\dagger)^* = (USS^\dagger U^*)^* = U(S^\dagger)^*S^*U^*$$
$$= U(SS^\dagger)^*U^* = USS^\dagger U^* = AA^\dagger.$$

It can be proved similarly that $(A^\dagger A)^* = A^\dagger A$.

---

[6]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).

(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 φ, $\varphi^\dagger$, and ψ 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 φ has the matrix S from the definition of the pseudoinverse $A^\dagger$, so we write directly

$$A = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \ A^\dagger = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

with the diagonal matrix D consisting of all non-zero singular values. We write again B for the matrix of ψ in these bases. Clearly B and A satisfy the assumptions of the claim (3). Thus

$$A^\dagger A = \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}, ABA = A$$

and we obtain

$$A^\dagger = A^\dagger ABAA^\dagger = \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} B \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

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 $\varphi : \mathbb{K}^n \to \mathbb{K}^m, x \mapsto Ax$, and direct sums $\mathbb{K}^n = (\mathrm{Ker}\,\varphi)^\perp \oplus \mathrm{Ker}\,\varphi$, $\mathbb{K}^m = \mathrm{Im}\,\varphi \oplus (\mathrm{Im}\,\varphi)^\perp$ of the orthogonal complements. The restricted mapping $\tilde{\varphi} := \varphi_{|(\mathrm{Ker}\,\varphi)^\perp} : (\mathrm{Ker}\,\varphi)^\perp \to \mathrm{Im}\,\varphi$ is a linear isomorphism. If we choose suitable orthonormal bases on $(\mathrm{Ker}\,\varphi)^\perp$ and $\mathrm{Im}\,\varphi$ and extend them to orthonormal bases on whole spaces, the mapping φ 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 $z \in \mathrm{Im}\,\varphi$ (in analytic geometry we shall say that the point z realises the distance of b from the affine subspace $\mathrm{Im}\,\varphi$), see 4.1.16). 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 $b = b_1 + b_2$, $b_1 \in \mathrm{Im}\,\varphi, b_2 \in (\mathrm{Im}\,\varphi)^\perp$.

Now, in our choice of bases, the mapping $\varphi^\dagger$ is given by the matrix $S^\dagger$ from the singular decomposition theorem. In

particular, $\varphi^\dagger(\operatorname{Im}\varphi) = (\operatorname{Ker}\varphi)^\perp$, D^{-1} is the matrix of the restriction $\varphi^\dagger_{|\operatorname{Im}\varphi}$, and $\varphi^\dagger_{|(\operatorname{Im}\varphi)^\perp}$ 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. $\qquad\square$

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 form 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, $j = 1, \ldots, m$. We want to find a linear combination of only a few fixed functions f_i, $i = 1, \ldots, n$ 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 ± 10.) This leads in our case to the matrix $A = (b_{ji})$

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 9 & 16 & 25 \end{pmatrix}.$$

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 $g(y) = x_1 + x_2 y + x_3 y^2$.
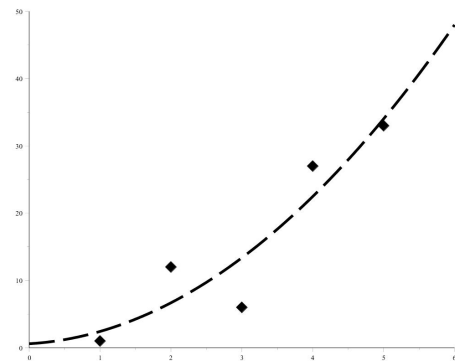
The computation was produced in Maple and taking 15 values $y_i = 1 + i + i^2$, with a random vector of deviations from the same range added produced the following picture:
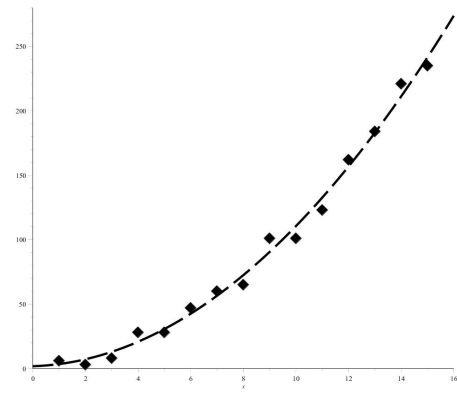
G. Additional exercises for the whole chapter

3.G.1. Solve the following LP problem

$$\text{minimize } \{7x - 5y + 3z\}$$

$$0 \le x \le 6, \quad -2 \le y \le 7, \quad -4 \le z \le 9.$$

Solution. Introduce new variables y' and z' by setting $y = y' - 2$, $z = z' - 4$. In the variables $\{x, y', z'\}$ the original LP problem takes the form

$$\text{maximize } \{-7x + 5y' - 3z'\}$$

$$x \le 6, \quad y' \le 9, \quad z' \le 13, \quad x, y', z' \ge 0.$$

Dropping the primes at y and z, we obtain an initial LP tableaux:

	x	y	z	t	s	r	
Objective	7	-5	3	0	0	0	0
t	1	0	0	1	0	0	6
s	0	①	0	0	1	0	9
r	0	0	1	0	0	1	13

The second and final tableaux, is

	x	y	z	t	s	r	
Objective	7	0	3	0	5	0	45
t	1	0	0	1	0	0	6
y	0	1	0	0	1	0	9
r	0	0	1	0	0	1	13

which provides the solution

$$x = 0, \; y' = 9, \; z' = 0.$$

In the original notation this means

$$x = 0, \; y = 7, \; z = -4.$$

This solution can be easily guessed from the beginning. $\qquad\square$

3.G.2. Solve the following LP problem: A small firm specializes in making five types of spare automotive parts. Each part is first cast from iron in the casting shop and is then sent to the finishing shop where holes are drilled, surfaces are turned, and edges are ground. The required worker-hours (per 100.units), for each type of parts in each of the two shops, are shown below:

Part type	1	2	3	4	5
Casting	2	1	3	3	1
Finishing	2	2	1	1	1

The profits from the five parts are \$30, \$20, \$40, \$25 and \$10 (per 100 units), respectively. The capacities of the casting and finishing shops over the next month are 700 and 1000 worker-hours respectively. Determine the quantities of each type of spare part to be made during the month, so as to maximize the firm's profit. Assume that there is sufficient demand for the firm to sell whatever it is capable of producing.

Solution. Let x_j be the number of produced units (in multiples of ten thousand) of the part of the type $j = 1, \ldots, 5$. Then the LP problem can be formulated as

$$\text{maximize}\,\{30x_1 + 20x_2 + 40x_3 + 25x_4 + 10x_5\}$$
$$2x_1 + x_2 + 3x_3 + 3x_4 + x_5 + t = 7$$
$$2x_1 + 2x_2 + x_3 + x_4 + x_5 + s = 10$$
$$x_j \geq 0, \ j = 1, \ldots, 5.$$

The first, second, and third LP tableaux are respectively

	x_1	x_2	x_3	x_4	x_5	t	s	
Objective	-3	-2	-4	$-\frac{5}{2}$	-1	0	0	0
t	②	1	3	3	1	1	0	7
s	2	2	1	1	1	0	1	10

	x_1	x_2	x_3	x_4	x_5	t	s	
Objective	0	$-\frac{1}{2}$	$\frac{1}{2}$	2	$\frac{1}{2}$	$\frac{3}{2}$	0	$\frac{21}{2}$
x_1	1	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{7}{2}$
s	0	①	-2	-2	0	-1	1	3

	x_1	x_2	x_3	x_4	x_5	t	s	
Objective	0	0	$-\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{24}{2}$
x_1	1	0	$\frac{5}{2}$	$\frac{5}{2}$	$\frac{1}{2}$	1	-1	2
x_2	0	①	-2	-2	0	-1	1	3

	x_1	x_2	x_3	x_4	x_5	t	s	
Objective	$\frac{1}{5}$	0	0	$\frac{3}{2}$	$\frac{3}{5}$	$\frac{6}{5}$	$\frac{3}{10}$	$\frac{62}{5}$
x_3	$\frac{2}{5}$	0	1	1	$\frac{1}{5}$	$\frac{2}{5}$	$-\frac{2}{5}$	$\frac{4}{5}$
x_2	$\frac{4}{5}$	1	0	0	$\frac{3}{5}$	$-\frac{1}{5}$	$\frac{1}{5}$	$\frac{23}{5}$

The final tableau provides the solution.

$$x_2 = \frac{23}{5}, \ x_3 = \frac{4}{5}, \ x_j = 0, \ j = 1, 4, 5.$$

Thus, optimal profit is achieved by producing $46,000$ parts of type 2 and $8,000$ parts of type 3. □

3.G.3. Model of spreading of annual plants. Consider some plants that blossom at the beginning of summer, then produce seeds and die at the peak of summer. Some of the seeds burst into flowers at the end of the autumn. Some survive the winter in the ground and burst into flowers at the start of the spring. The flowers that burst out in autumn and survive the winter are usually larger in the spring, and usually produce more seeds. After this, the whole cycle repeats itself.

The year is thus divided into four parts and in each of these parts we distinguish between some "forms" of the flower:

Part	Stage
beginning of spring	small and big seedlings
beginning of summer	small, medium and big blossoming flowers
peak of summer	seeds
autumn	seedlings and seeds

Denote by $x_1(t)$ and by $x_2(t)$ the number of small and large seedlings respectively at the start of the spring in year t. Denote by $y_1(t)$, $y_2(t)$ and $y_3(t)$ the number of small, medium and large flowers respectively in the summer of that year. From the

small seedlings either small or large flowers grow. From the large seedlings either medium or big flowers grow. Each of the seedlings can of course die (weather, be eaten by a cow, etc.) and nothing grows out of it. Denote by b_{ij} the probability that the seedling of the j-th size, $j = 1, 2$ grows into a flower of the i-th size, $i = 1, 2, 3$. Then we have

$$0 < b_{11} < 1, \quad b_{12} = 0, \quad 0 < b_{21} < 1, \quad 0 < b_{22} < 0,$$
$$b_{31} = 0, \quad 0 < b_{32} < 1, b_{11} + b_{21} < 1, \quad b_{22} + b_{32} < 1$$

(think in detail about what each of these inequalities expresses). If we consider classical probability, we can compute b_{11} as a ratio of the positive results (small seedling grows into a small flower) and of all possible results (the number of small seedlings). That is, $b_{11} = y_1(t)/x_1(t)$, or

$$y_1(t) = b_{11}x_1(t).$$

Analogously,

$$y_3(t) = b_{32}x_2(t).$$

Denote for a while by $y_{2,1}(t)$ and $y_{2,2}(t)$ the number of medium flowers that grow out of small and large seedlings respectively. Then $y_2(t) = y_{2,1}(t) + y_{2,2}(t)$ and $b_{21} = y_{2,1}(t)/x_1(t)$, $b_{22} = y_{2,2}(t)/x_2(t)$ and thus

$$y_2(t) = b_{21}x_1(t) + b_{22}x_2(t).$$

Write

$$B = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \\ 0 & b_{32} \end{pmatrix}, \ x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \ y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix}$$

and rewrite the previous equation in matrix notation

$$y(t) = Bx(t).$$

Denote by c_{11}, c_{12} and c_{13} the number of seeds produced by small, medium and large flowers respectively. Denote by $z(t)$ the total number of produced seeds in the summer of year t. Then

$$z(t) = c_{11}y_1(t) + c_{12}y_2(t) + c_{13}y_3(t).$$

In matrix calculus

$$z(t) = Cy(t)$$

with the notation

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} \end{pmatrix}.$$

If we want the matrix C to describe the modelled reality, we assume that the inequalities

$$0 < c_{11} < c_{12} < c_{13}$$

hold.

Finally, denote by $w_1(t)$ and $w_2(t)$ the number of seeds that burst in the autumn and the number of seeds that stay in the ground during the winter respectively. Denote by d_{11} and d_{21} the probabilities that the seeds burst out in the autumn and that the seeds do not burst respectively. Denote by f_{11} and f_{22} the probabilities that the seedling and the seed do not die during the winter respectively. The probabilities d_{11}, d_{21} must satisfy the inequalities

$$0 < d_{11}, \quad 0 < d_{21}, \quad d_{11} + d_{21} = 1.$$

Since a seedling dies in the winter more easily than a seed hidden in the ground, we assume that

$$0 < f_{11} < f_{22} < 1.$$

When denoting

$$D = \begin{pmatrix} d_{11} \\ d_{21} \end{pmatrix}, \qquad F = \begin{pmatrix} f_{11} & 0 \\ 0 & f_{22} \end{pmatrix}, \qquad w(t) = \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix}$$

we obtain, with similar ideas as before, the equalities

$$w(t) = Dz(t), \qquad x(t+1) = Fw(t).$$

Because the matrix multiplication is associative, we can compose the recurrent formulas:

$$
\begin{aligned}
x(t+1) = Fw(t) = F\big(Dz(t)\big) &= (FD)z(t) \\
&= (FD)\big(Cy(t)\big) = (FDC)y(t) \\
&= (FDC)\big(Bx(t)\big) = (FDCB)x(t), \\
y(t+1) = Bx(t+1) = B\big(Fw(t)\big) &= (BF)w(t) \\
&= (BF)\big(Dz(t)\big) = (BFD)z(t) = (BFD)\big(Cy(t)\big) \\
&= (BFDC)y(t), \\
z(t+1) = Cy(t+1) = C\big(Bx(t+1)\big) &= (CB)x(t+1) \\
&= (CB)\big(Fw(t)\big) = (CBF)w(t) = (CBF)\big(Dz(t)\big) \\
&= (CBFD)z(t), \\
w(t+1) = Dz(t+1) = D\big(Cy(t+1)\big) &= (DC)y(t+1) \\
&= (DC)\big(Bx(t+1)\big) = (DCB)x(t+1) \\
&= (DCB)\big(Fw(t)\big) = (DCBF)w(t).
\end{aligned}
$$

Using the notation

$$A_x = FDCB, \ A_y = BFDC, \ A_z = CBFD, \ A_w = DCBF,$$

we simplify them into the formula

$$x(t+1) = A_x x(t), \ y(t+1) = A_y y(t), \ z(t+1) = A_z z(t),$$
$$w(t+1) = A_w(t).$$

From these formulas we can compute the distribution of the population of the flowers in any part of any year, if we know the starting distribution of the population (that is, in the year zero).

For instance, let the distribution of the population be known in the summer, that is, $z(0)$ of seeds. The distribution of the population at the beginning of the spring in the t-th year is

$$
\begin{aligned}
x(t) &= A_x x(t-1) = A_x^2 x(t-2) = \cdots = A_x^{t-1} x(1) \\
&= A_x^{t-1} Fw(0) = A_x^{t-1} FDz(0).
\end{aligned}
$$

Note that the matrix $A_z = CBFD$ is of the type 1×1; it is not a matrix but just a scalar. We can denote by $\lambda = A_z$, and compute

(1) $\quad \lambda = CBFD$

$$= \begin{pmatrix} c_{11} & c_{12} & c_{13} \end{pmatrix} \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \\ 0 & b_{32} \end{pmatrix} \begin{pmatrix} f_{11} & 0 \\ 0 & f_{22} \end{pmatrix} \begin{pmatrix} d_{11} \\ d_{21} \end{pmatrix}$$

$$= \begin{pmatrix} c_{11}b_{11} + c_{12}b_{21} & c_{12}b_{22} + c_{13}b_{32} \end{pmatrix} \begin{pmatrix} f_{11}d_{11} \\ f_{22}d_{21} \end{pmatrix}$$

$$= b_{11}c_{11}d_{11}f_{11} + b_{21}c_{12}d_{11}f_{11} + b_{22}c_{12}d_{21}f_{22}$$

$$+ b_{32}c_{13}d_{21}f_{22}$$

and order the previous computation into a suitable form

$$x(t) = (FDCB)^{t-1}FDz(0) = FD(CBFD)^{t-2}CBFDz(0)$$

$$= FD(CBFD)^{t-1}z(0) = FDA_z^{t-1}z(0)$$

$$= \lambda^{t-1}FDz(0).$$

In this way only two matrix multiplications remain.

We list concrete values of the matrices B, C, D, F; they are the parameters of a hypothetical flower, which were inspired by the actual grass *Vulpia ciliata*:

$$B = \begin{pmatrix} 0.3 & 0 \\ 0.1 & 0.6 \\ 0 & 0.2 \end{pmatrix}, \ C = \begin{pmatrix} 1 & 10 & 100 \end{pmatrix}, \ D = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix},$$

$$F = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

Now we can compute the individual matrices, which map the vector describing the distribution of the population in some vegetative part of the year on the vector of the distribution of the population in the same part of the next year:

$$A_x = \begin{pmatrix} 0.0325 & 0.6500 \\ 0.0650 & 1.3000 \end{pmatrix} \quad A_y = \begin{pmatrix} 0.0075 & 0.0750 & 0.7500 \\ 0.0325 & 0.3250 & 3.2500 \\ 0.0100 & 0.1000 & 1.0000 \end{pmatrix},$$

$$A_z = 1.3325, \quad A_w = \begin{pmatrix} 0.0325 & 1.3000 \\ 0.0325 & 1.3000 \end{pmatrix}.$$

The value of $\lambda = A_z = 1.3325$ expresses the relative increment of the population between two years. Check for yourself that each of the matrices A_x, A_y, A_w has only one non-zero eigenvalue $\lambda = 1.3325$. The other eigenvalues are equal to 0.

We show one more application of the given model. We are interested in the "flexibility" of the reaction of the relative increment λ on the change of the individual "demographic parameters" – for instance, how the change of the probabilities of survival of the seeds changes the yearly increment. We reformulate formulate the question. By the *flexibility of the reaction of the characteristic λ on the parameter s*, denoted by $e(\lambda, s)$, we mean the relative change of the value λ related to the relative change of the parameter s. Even more precisely: by $\lambda(s)$ we denote the yearly increment in dependence on the parameter s. Then $\Delta\lambda(s) = \lambda(s + \Delta s) - \lambda(s)$ expresses the absolute change of the relative increment λ with the absolute change of the parameter s by Δs. The relative change of the increment of the parameter s is $\Delta s/s$. The flexibility is then the ratio of these two relative changes, that is,

$$e(\lambda, s) = \frac{\Delta\lambda(s)/\lambda(s)}{\Delta s/s} = \frac{s}{\lambda(s)} \frac{\lambda(s + \Delta s) - \lambda(s)}{\Delta s}.$$

Specifically, the yearly relative increment of the population depending on the survival of the seeds over the winter is, according to (1)

$$\lambda(f_{22}) = d_{21}(b_{22}c_{12} + b_{32}c_{13})f_{22} + d_{11}(b_{11}c_{11}f_{11} + b_{21}c_{12}f_{11})$$

and for specific values of the other parameters

$$\lambda(f_{22}) = 13f_{22} + 0.0325.$$

Because $f_{22} = 0.1$, we can compute

$$\lambda(0.1) = 1.3325, \quad \lambda(0.1 + \Delta s) = 1.3325 + 13\Delta s,$$

$$\Delta\lambda(0.1) = 13\Delta s,$$

therefore

$$e(\lambda, 0.1) = \frac{0.1}{1.3325} \frac{13\Delta s}{\Delta s} \doteq 0.976.$$

Analogically we can compute the flexibility of the reaction of the relative increment λ of the population on the other "demographic parameters". The results are summarised in the table

parameter	flexibility	parameter	flexibility
b_{11}	0.006	c_{11}	0.006
b_{21}	0.019	c_{12}	0.244
b_{22}	0.225	c_{13}	0.751
b_{23}	0.750	f_{11}	0.024
d_{11}	0.024	f_{22}	0.976
d_{21}	0.976		

From it we can see that the increment λ is mostly influenced by the number of the seeds that overwinter (parameter d_{21}) and their survivability (parameter f_{22}). This revelation is not surprising. Farmers have been aware of this fact since neolithic times. The result shows that the mathematical model adequately describes the reality.

3.G.4. 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 action of the farmer) is

$$L = \begin{pmatrix} 0 & 2 & 5 & 2 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

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 we are dealing with the model

$$L = \begin{pmatrix} 0 & 2 & 5 & 2 \\ a & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

We are looking for an 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 it to have 1 as a root, then $a = \frac{1}{5}$ The farmer can thus 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 $(20, 4, 2, 1)$ and in these ratios the population stabilises.

$\square$

3.G.5. 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 exists no rat 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). $\bigcirc$

3.G.6. Model of evolution of a whale population. For the evolution of a population, females are important. The important factor is not age but fertility. From this point of view we 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 postclimacterial state depends 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. A young fertile female has less children per unit interval than an adult female. We 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, and by p_2 and p_3 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 the probabilities respectively as q_2, q_3 and q_4 for young, adult and old females. Each of the numbers $p_1, p_2, p_3, q_2, q_3, q_4$ is a probability from the interval $[0.1]$.

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

$$x_1(t+1) = f_2 x_2(t) + f_3 x_3(t).$$

Denote temporarily by $x_{2,1}(t+1)$ the number of young females in time $t+1$, who were juvenile in the previous time interval.

Denote temporarily by $x_{2,2}(t+1)$ the number of young females, who were already fertile in time t, survived that time interval, bud 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 $x_{2,2}(t+1)/x_2(t)$. Since young females in time $t+1$ are exactly those who survived the juvenile stage and were already fertile, did survive and did not evolve,

$$x_2(t+1) = x_{2,1}(t+1) + x_{2,2}(t+1) = p_1 x_1(t) + q_2 x_2(t).$$

Similarly, the expected number of fully fertile females is

$$x_3(t+1) = p_2 x_2(t) + q_3 x_3(t)$$

and the expected number of postclimacterial females is

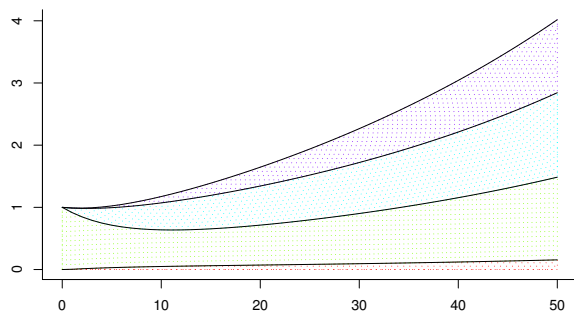$$x_4(t+1) = p_3 x_3(t) + q_4 x_4(t).$$

FIGURE 1. Evolution of a population of orca whale. On the horizontal axis the time is in years, on the vertical axis is the size of the population. Individual areas depict the number of juvenile, young, adult and old females respectively, from below.

Now we can denote

$$A = \begin{pmatrix} 0 & f_2 & f_3 & 0 \\ p_1 & q_2 & 0 & 0 \\ 0 & p_2 & q_3 & 0 \\ 0 & 0 & p_3 & q_4 \end{pmatrix}, \qquad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{pmatrix}$$

and rewrite the previous recurrent formulas in matrix form

$$x(t+1) = Ax(t).$$

Using this matrix difference equation we can compute the expected number of females in individual categories, if we know the distribution of the population at some initial time.

Specifically, for the population of orca whales the following parameters were observed:
$$p_1 = 0.9775, \quad q_2 = 0.9111, \quad f_2 = 0.0043,$$
$$p_2 = 0.0736, \quad q_3 = 0.9534. \quad f_3 = 0.1132,$$
$$p_3 = 0.0452, \quad q_4 = 0.9804;$$
The time interval is in this case 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 $x(0) = (0,1,0,0)^T$, we can compute

$$x(1) = \begin{pmatrix} 0 & 0.0043 & 0.1132 & 0 \\ 0.9775 & 0.9111 & 0 & 0 \\ 0 & 0.0736 & 0.9534 & 0 \\ 0 & 0 & 0.0452 & 0.9804 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.0043 \\ 0.9111 \\ 0.0736 \\ 0 \end{pmatrix},$$

$$x(2) = \begin{pmatrix} 0 & 0.0043 & 0.1132 & 0 \\ 0.9775 & 0.9111 & 0 & 0 \\ 0 & 0.0736 & 0.9534 & 0 \\ 0 & 0 & 0.0452 & 0.9804 \end{pmatrix} \begin{pmatrix} 0.0043 \\ 0.9111 \\ 0.0736 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.01224925 \\ 0.83430646 \\ 0.13722720 \\ 0.00332672 \end{pmatrix}$$

The results of the computation can be also expressed graphically; see the diagram 1. 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.

The matrix A thus has the eigenvalues

$\lambda_1 = 1.025441326, \lambda_2 = 0.980400000, \lambda_3 = 0.834222976, \lambda_4 = 0.004835698.$

The eigenvector associated with the largest eigenvalue λ_1 is

$w = (0.03697187, 0.31607121, 0.32290968, 0.32404724);$

210

this vector is normed such that the sum of its components equals 1.

Compare the evolution of the size of the population with the exponential function $F(t) = \lambda_1^t x_0$, where x_0 is the total size of the initial population. Compute also the relative distribution of individual categories in the population after a certain time of evolution. Compare it with the components of the eigenvector w. They will appear very close, this is caused by the fact that A has only a single eigenvalue with the greatest absolute value. Also, the vector space generated by the eigenvectors associated with the eigenvalues $\lambda_2, \lambda_3, \lambda_4$ has only the zero vector with the non-negative orthant intersection. The structure of the matrix A itself does not ensure such an easily predictable evolution, since it is a reducible matrix (see ??).

3.G.7. Model of growth of population of teasels *Dipsacus sylvestris*. This plant can be seen in four stages. Either as a blossoming plant or as a rosette of leaves. With the rosette there are three sizes – small, medium and large. The life cycle of this monoecious perennial plant can be described as follows.

A blossoming plant produces a number of seeds in late summer and dies. From the seeds, some of them sprout in that year into a rosette of leaves, usually of medium size. Other seeds spend the winter in the ground. Some of the seeds in the ground sprout in the spring into a rosette, but because they were weakened during the winter, the size is usually small. After three or more winters the "sleeping" (formally, dormant) seeds die as they loose the ability to sprout. Depending on the environment of the plant, a small or medium rosette can grow during the year, and any rosette can stay in its category or die (wither, be eaten by insects, etc.) A medium or large rosette can burst into a flower in the next year. A blossoming flower then produces seeds and the cycle repeats.

In order to be able to predict the spreading of the population of the teasels, we need to quantify the described events. Botanists discovered that a blossoming plant produces on average 431 seeds. The probabilities that a seed sprouts, that a rosette grows or bursts into a flower are summarised in the following table:

event	probability
seed produced by a flower dies	0.172
seed sprouts into a small rosette in the current year	0.008
seed sprouts into a medium rosette in the current year	0.070
seed sprouts into a large rosette in the current year	0.002
seed sprouts into a small rosette after spending the winter	0.013
seed sprouts into a medium rosette after spending the winter	0.007
seed sprouts into a large rosette after spending the winter	0.001
seed sprouts into a small rosette after spending two winters	0.001
seed dies after spending one winter	0.013
small rosette survives but does not grow	0.125
medium rosette survives but does not grow	0.238
large rosette survives but does not grow	0.167
small rosette grows into a medium one	0.125
small rosette grows into a large one	0.036
medium rosette grows into a large one	0.245
medium rosette bursts into a flower	0.023
large rosette bursts into a flower	0.750

Note that all the relevant events in the life cycle have their probabilities given and that the events are mutually incompatible.

Imagine that we always observe the population at the beginning of the vegetative year, say in March, and that all considered events take place in the rest of the year, say from April to February. In the population there are blossoming flowers, rosettes of three sizes, produced seeds and seeds that have been dormant for a year or two. This leads to a division of the population into seven classes – just-produced seeds, seeds dormant for one year, seeds dormant for two years, rosettes small, medium and large and blossoming flowers. But the just-produced seeds are changed either into rosettes or they spend winter in the same year, thus they do not form an individual category. We denote:

$x_1(t)$ — the number of seeds dormant for one year in the spring of the year t
$x_2(t)$ — the number of seeds dormant for two years in the spring of the year t
$x_3(t)$ — the number of small rosettes in the spring of the year t
$x_4(t)$ — the number of medium rosettes in the spring of the year t
$x_5(t)$ — the number of large rosettes in the spring of the year t
$x_6(t)$ — the number of blossoming flowers in the spring of the year t

The number of produced seeds in the year t is $431x_6(t)$. The probability that a seed stays dormant for the first year equals the probability that the seed does not sprout into a rosette and does not die, that is, $1 - (0.008 + 0{,}070 + 0.002 + 0{,}172) = 0.748$. The expected number of seeds dormant for winter in the next year is thus

$$x_1(t+1) = 0.748 \cdot 431x_6(t) = 322.388x_6(t).$$

The probability that the seed that has been dormant for one year stays dormant for the second year equals the probability that the dormant seed does not sprout into a rosette and that it does not die, that is, $1 - 0.013 - 0.007 - 0.001 - 0.013 = 0.966$. The expected number of seeds dormant for two winters is thus

$$x_2(t+1) = 0.966x_1(t).$$

A small rosette can sprout from the seeds immediately, from a seed dormant for one year or from a seed dormant for two years. The expected number of small rosettes sprouted from non-dormant seeds in the year t equals $0.008 \cdot 431x_6(t) = 3.448x_6(t)$. The expected number of small rosettes sprouted from the seeds dormant for one and two years is $0.013x_1(t)$ and $0.010x_2(t)$ respectively. With these newly sprouted small rosettes there are in the population also the older small rosettes (those that have not grown yet) – of those there are $0.125x_3(t)$. The total expected number of small rosettes is thus

$$x_3(t+1) = 0.013x_1(t) + 0.010x_2(t) + 0.125x_3(t) + 3.448x_6(t).$$

Analogically we determine the expected number of medium and large rosettes

$$x_4(t+1) = 0.007x_1(t) + 0.125x_3(t) + 0.238x_4(t) + 0.070 \cdot 431x_6(t) =$$
$$= 0.007x_1(t) + 0.125x_3(t) + 0.238x_4(t) + 30.170x_6,$$

$$x_5(t+1) = 0.245x_4(t) + 0.167x_5(t) + 0.002 \cdot 431x_6(t) =$$
$$= 0.245x_4(t) + 0.167x_5(t) + 0.862x_6(t).$$

The blossoming flower can arise either from medium or from large rosette. The expected number of blossoming flowers is thus

$$x_6(t+1) = 0{,}023x_4(t) + 0{,}750x_5(t).$$

We have thus reached six recurrent formulas for individual components of the investigated plant. We now denote

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 322.388 \\ 0.966 & 0 & 0 & 0 & 0 & 0 \\ 0.013 & 0{,}010 & 0.125 & 0 & 0 & 3.448 \\ 0.007 & 0 & 0.125 & 0.238 & 0 & 30.170 \\ 0.008 & 0 & 0.038 & 0.245 & 0.167 & 0.862 \\ 0 & 0 & 0 & 0.023 & 0.750 & 0 \end{pmatrix}, \quad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \end{pmatrix}$$

and write the previous equalities in matrix form suitable for the computation

$$x(t+1) = Ax(t).$$

If we know the distribution of the individual components of the population in some initial year $t = 0$, then we can compute the expected numbers of flowers and seeds in the following years. We can also compute the total number of individuals $n(t)$ at the time t, $n(t) = \sum_{i=1}^{6} x_i(t)$. We can compute the relative distribution of the individual components $x_i(t)/n(t)$, $i = 1, 2, 3, 4, 5, 6$ and the yearly relative change in the population $n(t+1)/n(t)$. The results of such calculations for fifteen years, and the case above of one blossoming flower, are given in the table 1. Unlike the whale population, the image would not be very clear, as

t	x_1	x_2	x_3	x_4	x_5	x_6	$n(t)$
0	0,00	0,00	0,00	0,00	0,00	1,00	1,00
1	322,39	0,00	3,45	30,17	0,86	0,00	356,87
2	0,00	311,43	4,62	9,87	10,25	1,34	337,50
3	432,13	0,00	8,31	43,37	5,46	7,91	497,18
4	2 550,50	417,44	33,93	253,07	22,13	5,09	3 282,16
5	1 641,69	2 463,78	59,13	235,96	91,78	22,42	4 514,76
6	7 227,10	1 585,88	130,67	751,37	107,84	74,26	9 877,12
7	23 941,29	6 981,37	382,20	2 486,25	328,89	98,16	34 218,17
8	31 646,56	23 127,29	767,29	3 768,67	954,73	303,85	60 568,39
9	97 958,56	30 570,58	1 786,27	10 381,63	1 627,01	802,72	143 126,78
10	258 788,42	94 627,97	4 570,24	27 597,99	4 358,70	1 459,04	391 402,36
11	470 376,19	249 989,61	9 912,57	52 970,28	10 991,08	3 903,78	798 143,52
12	1 258 532,41	454 383,40	23 314,10	134 915,73	22 317,98	9 461,62	1 902 925,24
13	3 050 314,29	1 215 742,31	56 442,70	329 291,15	55 891,57	19 841,54	4 727 523,56
14	6 396 675,73	2 946 603,60	127 280,49	705 398,22	133 660,97	49 492,37	10 359 111,38
15	15 955 747,76	6 179 188,75	299 182,59	1 721 756,52	293 816,44	116 469,89	24 566 161,94

t	$\dfrac{x_1(t)}{n(t)}$	$\dfrac{x_2(t)}{n(t)}$	$\dfrac{x_3(t)}{n(t)}$	$\dfrac{x_4(t)}{n(t)}$	$\dfrac{x_5(t)}{n(t)}$	$\dfrac{x_6(t)}{n(t)}$	$\dfrac{n(t+1)}{n(t)}$
0	0,000	0,000	0,000	0,000	0,000	1,000	356,868
1	0,903	0,000	0,010	0,085	0,002	0,000	0,946
2	0,000	0,923	0,014	0,029	0,030	0,004	1,473
3	0,869	0,000	0,017	0,087	0,011	0,016	6,602
4	0,777	0,127	0,010	0,077	0,007	0,002	1,376
5	0,364	0,546	0,013	0,052	0,020	0,005	2,188
6	0,732	0,161	0,013	0,076	0,011	0,008	3,464
7	0,700	0,204	0,011	0,073	0,010	0,003	1,770
8	0,522	0,382	0,013	0,062	0,016	0,005	2,363
9	0,684	0,214	0,012	0,073	0,011	0,006	2,735
10	0,661	0,242	0,012	0,071	0,011	0,004	2,039
11	0,589	0,313	0,012	0,066	0,014	0,005	2,384
12	0,661	0,239	0,012	0,071	0,012	0,005	2,484
13	0,645	0,257	0,012	0,070	0,012	0,004	2,191
14	0,617	0,284	0,012	0,068	0,013	0,005	2,371
15	0,650	0,252	0,012	0,070	0,012	0,005	

TABLE 1. Modelled evolution of the population of teasels *Dipsacus sylvestris*. Sizes of the individual components of population, the total size of population, relative distribution of the individual components of population and the relative increments of sizes.

the numbers of flowers are negligible compared to the numbers of seeds (the individual areas for flowers would merge in the picture).

The matrix A has eigenvalues

$\lambda_1 = 2.3339$ $\lambda_4 = 0.1187 + 0.1953i$
$\lambda_2 = -0.9569 + 1.4942i$ $\lambda_5 = 0.1187 - 0.1953i$
$\lambda_3 = -0.9569 - 1.4942i$ $\lambda_6 = -0.1274$

The eigenvector associated with the eigenvalue λ_1 is

$w = (0.6377,\ 0.2640,\ 0.0122,\ 0.0693,\ 0.0122,\ 0.0046);$

this vector is normed such that the sum of its components is one. With increasing time t, the relative increment in the size of the population approaches the eigenvalue λ_1, the relative distribution of the components in the population approach the components of the normed eigenvector associated with the eigenvector λ_1. Every non-negative matrix that has non-zero elements in the same positions as A is primitive. The evolution of the population necessarily approaches a stable structure.

213

3.G.8. Nonlinear model of population. Investigate in detail the evolution of the population for a non-linear model from the text book (1.12), with $K = 1$ and

  i) rate of growth $r = 1$ and the initial state $p(1) = 0.2$
 ii) rate of growth $r = 1$ and the initial state $p(1) = 2$
iii) rate of growth $r = 1$ and the initial state $p(1) = 3$
 iv) rate of growth $r = 2.2$ and the initial state $p(1) = 0.2$
  v) rate of growth $r = 3$ and the initial state $p(1) = 0.2$

Compute some members of the sequence and predict the future growth of the population.

Solution.

  i) The first ten members of the sequence $p(n)$ is in the following table. From there we can see that the size of the population converges to the value 1.

n	p(n)
1	0.2
2	0.36
3	0.5904
4	0.83222784
5	0.971852502
6	0.999207718
7	0.999999372

Graph for the evolution of the population for $r = 1$ and $p(1) = 0.2$:

 ii) For the initial value $p(1) = 2$ we obtain $p(2) = 0$ and after that the population does not change.

iii) For $p(1) = 3$ we obtain

n	p(n)
1	3
2	-15
3	-255
4	-65535

and from there we see that the populations decreases under all bounds.

 iv) For the measure of growth $r = 2,2$ and the initial state $p(1) = 0.2$ we obtain

n	p(n)
1	0.2
2	0.552
3	1.0960512
4	0,864441727
5	1.122242628
6	0.820433675
7	1.144542647
8	0.780585155
9	1.157383491
10	0.756646772
11	1.161738128
12	0.748363958
!3	1.162657716
14	0.74660417

Instead of convergence we obtain in this case an oscillation – after some time the population jumps between the values 1,16 and 0.74. The graph of the evolution of the population for $r = 2,2$ and $p(1) = 0.2$ then looks as follows:

  v) For the rate of growth $r = 3$ and the initial state $p(1) = 0.2$ we obtain

214

n	p(n)
1	0.2
2	0.68
3	1.3328
4	0.00213248
5	0.008516278
6	0.033847529
7	0.131953152
8	0.475577705
9	1.223788359
10	0.402179593
11	1.123473097
12	0.707316989
13	1.328375987
14	0.019755658
15	0.077851775
16	0.293224403
17	0.91495596
18	1.148390614
19	0.63715945
20	1.330721306
21	0.010427642
22	0.041384361
23	0.160399447

In this case the situation is more complicated – the population starts oscillating between more values. In order to be able to see between what values, we would need to compute more members. For the members from the table we have the following graph:

□

3.G.9. 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 only 0.6.

This process is continued indefinitely. For any $n \in \mathbb{N}$ determine the probability that the n-th experiment is successful.

Solution. Introduce the probabilistic vector

$$x_n = \left(x_n^1, x_n^2\right)^T, \quad n \in \mathbb{N},$$

where x_n^1 is the probability of the success of the n-th experiment and $x_n^2 = 1 - x_n^1$ is the probability of its failure. According to the statement

$$x_1 = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

and hence also

$$x_2 = \begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix} \cdot \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 13/20 \\ 7/20 \end{pmatrix}.$$

Using the notation

$$T = \begin{pmatrix} 7/10 & 3/5 \\ 3/10 & 2/5 \end{pmatrix}$$

it holds that

(1) $$x_{n+1} = T \cdot x_n, \quad n \in \mathbb{N},$$

because the probabilistic vector x_{n+1} depends only on x_n and this dependency is identical for both x_2 and x_1. From the relation (1) we have directly

(2) $x_{n+1} = T \cdot T \cdot x_{n-1} = \cdots = T^n \cdot x_1, \quad n \geq 2, n \in \mathbb{N}.$

Therefore we express T^n, $n \in \mathbb{N}$. It is a Markov process, and thus 1 is an eigenvalue of the matrix T. The second eigenvalue 0.1, and follows for instance from the fact that the trace (the sum of the elements on the diagonal) equals the sum of the eigenvalues (every eigenvalue is counted with its algebraic multiplicity). To these eigenvalues then correspond the eigenvectors

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

We thus obtain

$$T = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1/10 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}^{-1},$$

that is, for $n \in \mathbb{N}$ we have

$$T^n = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1/10 \end{pmatrix}^n \cdot \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}^{-1} =$$

$$= \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1^n & 0 \\ 0 & 10^{-n} \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}^{-1}.$$

Substitution

$$\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}^{-1} = \frac{1}{3} \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$$

and multiplication yields

$$T^n = \frac{1}{3} \begin{pmatrix} 2 + 10^{-n} & 2 - 2 \cdot 10^{-n} \\ 1 - 10^{-n} & 1 + 2 \cdot 10^{-n} \end{pmatrix}, \quad n \in \mathbb{N}.$$

From there, from (1) and from (2) it follows that

$$x_{n+1} = \left(\frac{2}{3} - \frac{1}{6 \cdot 10^n}, \frac{1}{3} + \frac{1}{6 \cdot 10^n} \right)^T, \quad n \in \mathbb{N}.$$

Specially, we see that for big n the probability of success of the n-th experiment is close to $2/3$. $\qquad \square$

3.G.10. A student in a student dormitory is very "socially tired". As a result, he is not able to fully perceive the universe around him and coordinate his movements. In this state he decides to invite his friend who lives at the end of the hall to the party-in-progress. But at the other end of the hall there lives somebody he definitely does not wish to invite.

He is so "tired", that he attains the decision to make a step in a desired direction only in 53 of 100 attempts (in the remaining 47, he makes a step in exactly the opposite direction).

Assuming that he starts in the middle of the hall and that the distance to both of the doors at the ends corresponds to twenty of his awkward steps, determine the probability that he first reaches the desired door.

$\bigcirc$

3.G.11. Let $n \in \mathbb{N}$ of persons be playing the "silent post". For simplicity, assume that the first person whispers to the second person exactly one (arbitrarily chosen) of the words "yes", "no". The second person then whispers to the third person

the choice of the words "yes", "no" that the second person thinks the first person whispered. This continues to the n-th person. If the probability that the word changes (on purpose or accidentally) to the alternative word during one transmission is $p \in (0.1)$, determine for large $n \in \mathbb{N}$ the probability that the n-th person correctly receives the same word as transmitted by the first person.

Solution. We can view this problem as a Markov chain with two states called Yes and No. We say that the process is in the "yes" state in time $m \in \mathbb{N}$, if the m-th person thinks that the received word is "yes". For the order of the states "yes", "no", the probabilistic matrix is

$$T = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

The product of the matrix T^{m-1} and the probabilistic vector of the initial choice of the first person then gives the probability of what the m-th person thinks. We do not have to compute the powers of this matrix, because all the elements of the matrix T are positive numbers. Furthermore, this matrix is doubly stochastic. Thus for large $n \in \mathbb{N}$ the probabilistic vector is close to the vector $(1/2, 1/2)^T$. The probability that the n-th person says "yes", is thus approximately the same as the probability that the n-th person says "no", independently of the initial word. For a large number of participants roughly half of them hears "yes". We repeat that this does not depend on the initial word.

For completeness, we determine what would be the result if we assumed that the probability of change from "yes", to "no" is for any person equal to $p \in (0.1)$ and the probability of change from "no" to "yes", equals (generally distinct) $q \in (0.1)$. In this case for the same order of the states we obtain a probabilistic matrix

$$T = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}.$$

This leads (for large $n \in \mathbb{N}$) to the probabilistic vector close to the vector

$$\left(\frac{q}{p+q}, \frac{p}{p+q} \right)^T,$$

which for instance follows from the expression of the matrix

$$T^n = \frac{1}{p+q} \left[\begin{pmatrix} q & q \\ p & p \end{pmatrix} + (1-p-q)^n \begin{pmatrix} p & -q \\ -p & q \end{pmatrix} \right].$$

Again, with sufficiently many people it does not depend on the initial choice of the word. Simply speaking, in this model, it does not depend on the initial state, because the people decide what the transmitted information is about; more precisely, the people themselves decide about the frequency of appearance of "yes" and "no", if there are enough of them and there is no checking present.

The obtained result was experimentally confirmed. In a psychological experiment, an individual was repeatedly exposed to an event that could have been interpreted in two ways. It was being done in time intervals that ensured that the subject still remembered the previous event. See for instance "T. Havránek et al.: *Matematika pro biologické a lékařské vědy*, Praha, Academia 1981", where there is an experiment in which an ambiguous object (say, a drawing of a cube which can be perceived from both the bottom and the top) is in fixed time intervals lighted on. Such process is a Markov chain with the transition matrix

$$\begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix},$$

where $p, q \in (0.1)$. $\qquad\square$

3.G.12. Petr regularly meets his friend. But he is "well-known" for his bad timekeeping. But he is trying to change. Thus in half of the cases he arrives on time, and in one tenth of the cases he comes even sooner, given that he was late for the previous meeting. But if he was on time or sooner for the last meeting, he returns back to his "carelessness" and with probability 0.8 he arrives late, and with only 0.2 he is on time. What is the probability that on the 20th meeting he arrives late, given that he was on time on the eleventh?

Solution. This is a Markov process with states "Petr comes late", "Petr comes on time", "Petr comes sooner" with the probabilistic transition matrix (with the given order of states)

$$T = \begin{pmatrix} 0.4 & 0.8 & 0.8 \\ 0.5 & 0.2 & 0.2 \\ 0.1 & 0 & 0 \end{pmatrix}.$$

The eleventh meeting is determined by the probabilistic vector $(0, 1, 0)^T$ (when Petr comes on time). To the twentieth meeting corresponds the vector

$$T^9 \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0,571\,578\,368 \\ 0,371\,316\,224 \\ 0,057\,105\,408 \end{pmatrix}.$$

The desired probability is thus $0,571\,578\,368$ (exactly). We add that

$$T^9 = \begin{pmatrix} 0.571\,316\,224 & 0.571\,578\,368 & 0.571\,578\,368 \\ 0.371\,512\,832 & 0.371\,316\,224 & 0.371\,316\,224 \\ 0.057\,170\,944 & 0.057\,105\,408 & 0.057\,105\,408 \end{pmatrix}.$$

From this, it is seen that it really does not depend on whether Petr came on the eleventh meeting late (first column), on time (second) or sooner (third). □

3.G.13. Two students A and B spend every Monday morning by playing a certain computer game. The person who wins then 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. If student B won a game, then with probability $1/2$ he 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 was won by A.

Solution. This a Markov process with the states "student A wins", "the game ends with a draw", "student B wins" (in this order), with the probabilistic transition matrix

$$T = \begin{pmatrix} 3/4 & 0 & 1/4 \\ 1/4 & 2/3 & 1/4 \\ 0 & 1/3 & 1/2 \end{pmatrix}.$$

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). 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_∞ of the matrix T associated with the eigenvalue 1. It is straightforward to compute

$$x_\infty = \left(\frac{2}{7}, \frac{3}{7}, \frac{2}{7} \right)^T.$$

The vector x_∞ differs only very slightly from the probabilistic vector for large n. It does not depend on the initial state. For large $n \in \mathbb{N}$, we obtain

$$T^n \approx \begin{pmatrix} 2/7 & 2/7 & 2/7 \\ 3/7 & 3/7 & 3/7 \\ 2/7 & 2/7 & 2/7 \end{pmatrix}.$$

The desired probability is the element of this matrix on the second position in the first column (the second component of the vector x_∞). Hence the result is $3/7$. □

3.G.14. Popularity of the media. 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, $1/5$ viewers of the second move to the first channel.

Determine the time evolution of the number of viewers watching given channels using Markov processes. Write down a matrix of the process, and find its eigenvalues and eigenvectors. ◯

3.G.15. Students at the lecture. Students can be divided into, say, three groups – 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 it is as follows:

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. 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. A student who is in the pub has 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. The matrix of the Markov process is $\begin{pmatrix} 0.5 & 0.1 & 0 \\ 0.4 & 0.5 & 0 \\ 0.1 & 0.4 & 1 \end{pmatrix}$. Its characteristic polynomial is $(0.5 - \lambda)^2(1 - \lambda) - 0.4(1 - \lambda) = 0$. Evidently one is an eigenvalue of this matrix (the other roots are 0.3 and 0.7). In the course of time, the students divide into groups as described by the corresponding eigenvector – which is a solution of the equality $\begin{pmatrix} -0.5 & 0.1 & 0 \\ 0.4 & -0.5 & 0 \\ 0.1 & 0.4 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0$. These are multiples of the vector $(0, 0, 1)$. In other words, all the students end up in the pub.

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. Adding 10 percent possibility for leaving the pub, this changes. The corresponding matrix is now $\begin{pmatrix} 0.5 & 0.1 & 0 \\ 0.4 & 0.5 & 0.1 \\ 0.1 & 0.4 & 0.9 \end{pmatrix}$. Again the state stabilises on the eigenvector associated with the eigenvalue 1. In this case the solution of the equation

$$\begin{pmatrix} -0.5 & 0.1 & 0 \\ 0.4 & -0.5 & 0.1 \\ 0.1 & 0.4 & -0.1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0$$

is wanted. A solution is for instance the vector $(1, 5, 21)$. The distribution of the students in the individual group is then given by the multiple of this vector such that the coordinates sum to one, that is, the vector $(\frac{1}{27}, \frac{5}{27}, \frac{21}{27})$. Again, most of the students end up in the pub, but some will be at school. □

3.G.16. Roulette. A roulette player has the following strategy: he comes to play with €10. He always bets everything he has. He always bets on black (there are 37 numbers in the roulette, 18 black, 18 red and zero). The player ends whenever he has either nothing, or when he wins €80. Consider this problem as a Markov process and write down its 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 its states, and we construct the matrix:

$$A = \begin{pmatrix} 1 & a & a & a & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & b & 0 & 0 & 0 \\ 0 & 0 & b & 0 & 0 \\ 0 & 0 & 0 & b & 1 \end{pmatrix},$$

where $a = \frac{19}{37}$ and $b = \frac{18}{37}$. Note that the matrix is probabilistic and singular. The eigenvalue 1 is a double one. The game does not converge to a single vector x_∞, but ends in one of the eigenvectors associated with eigenvalue 1, that is, either $(1, 0, 0, 0, 0)$ (the player looses it all), or $(0, 0, 0, 0, 1)$ (the player wins €80). Furthermore we observe that the game ends

after three bets, that is, the sequence $\{A^n\}_{n=1}^{\infty}$, is constant for $n \geq 3$:

$$A^{\infty} := A^3 = A^n = \begin{pmatrix} 1 & a+ab+ab^2 & a+ab & a & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & b^3 & b^2 & b & 1. \end{pmatrix}$$

We easily determine 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. (We multiply by the matrix A^{∞} the initial vector $(0, 1, 0, 0, 0)$ and obtain the vector $(a + ab + ab^2, 0, 0, 0, b^3)$.) □

3.G.17. 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}. ○

3.G.18. Based on the temperature at 14:00, the days are divided into warm, average, and cold. From the all-year statistics, after a warm day the next day is warm in 50 % of the cases and is average in 30 % of the cases, after an average day the next day is average in 40 % of the cases, and cold in 30 % of the cases, and after a cold day, the next day is cold in 50 % of the cases, and average in 30 % of the cases.

Without any further information, derive how many warm, cold, and average days can be expected in a year.

Solution. For each day exactly one of the states warm day, average day, cold day is attained. If the vector x_n has as its components the probabilities that a certain (n-th) day is warm, average and cold (respectively), then the components of the vector

$$x_{n+1} = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \cdot x_n$$

show the probabilities that the next day is warm, average and cold respectively. To verify, it suffices to substitute

$$x_n = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad x_n = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad x_n = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

while for instance for the third choice we must obtain the probabilities that after a cold day there follows a warm, average and cold day respectively. We see that the problem is a Markov chain problem with probabilistic transitional matrix

$$T = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}.$$

Because all the elements of this matrix are positive, there exists a probabilistic vector

$$x_{\infty} = \left(x_{\infty}^1, x_{\infty}^2, x_{\infty}^3\right)^T,$$

to which the vector x_n approaches as n grows, independently of the vector x_n for small n. Furthermore, by the corollary of the Perron-Frobenius theorem, x_{∞} is the eigenvector of the matrix T with the eigenvalue 1. Thus

$$\begin{array}{rclclcl} x_{\infty}^1 & = & 0.5\,x_{\infty}^1 & + & 0.3\,x_{\infty}^2 & + & 0.2\,x_{\infty}^3, \\ x_{\infty}^2 & = & 0.3\,x_{\infty}^1 & + & 0.4\,x_{\infty}^2 & + & 0.3\,x_{\infty}^3, \\ x_{\infty}^3 & = & 0.2\,x_{\infty}^1 & + & 0.3\,x_{\infty}^2 & + & 0.5\,x_{\infty}^3, \\ 1 & = & x_{\infty}^1 & + & x_{\infty}^2 & + & x_{\infty}^3, \end{array}$$

where the last condition means that the vector x_{∞} is probabilistic. It is easy to see that this system has a unique solution

$$x_{\infty}^1 = x_{\infty}^2 = x_{\infty}^3 = \frac{1}{3}.$$

Thus we can expect roughly the same number of warm, average and cold days.

We emphasise that the sum of the numbers from any column of the matrix T must equal 1,otherwise it would not be a Markov process. Because $T^T = T$, the matrix T is symmetric, and the sum of all numbers from any row also equals 1. We say that a matrix with non-negative elements and with the property that the sum of the numbers in any column or in any row

equals one is called doubly stochastic. An important property of every doubly stochastic primitive matrix (for any dimension – the number of states) is that the corresponding vector x_∞ has all of its components identical. That is, after sufficiently many iterations all the states in the corresponding Markov chain are attained with the same frequency. □

3.G.19. John 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. Assume that he has been running like this for very many days. How often does he choose the short track and how often the long track? What is the probability that he chooses a long track when he picked it a week before?

Solution. Clearly it is a Markov process with three possible states – choices for a short, medium or long track. This order of the states gives a probabilistic transition matrix

$$T = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{pmatrix}.$$

It suffices to observe that (for instance) the second column 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). Since

$$T^2 = \begin{pmatrix} 1/6 & 1/6 & 1/9 \\ 5/12 & 5/12 & 4/9 \\ 5/12 & 5/12 & 4/9 \end{pmatrix},$$

we can use the corollary of the Perron-Frobenius theorem for Markov chains. It is not difficult to compute the eigenvector corresponding to the eigenvalue 1. It is a probabilistic vector, namely:

$$\left(\frac{1}{7}, \frac{3}{7}, \frac{3}{7} \right)^T.$$

The numbers $1/7, 3/7, 3/7$ are then respectively the probabilities that in a randomly chosen day he choose a short, medium or long track.

Suppose on a certain day, John (that is, in time $n \in \mathbb{N}$) chooses a long track. This corresponds to the probabilistic vector

$$x_n = (0, 0, 1)^T.$$

For the following day,

$$x_{n+1} = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix},$$

and after seven days

$$x_{n+7} = T^7 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = T^6 \cdot \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

The enumeration gives us as components of x_{n+7} the values

$0.142\,861\,225\ldots;\quad 0.428\,569\,387\ldots;\quad 0.428\,569\,387\ldots$

Thus the probability that he chooses a long track under the condition that he chose it seven days ago is roughly $0.428\,569 \approx 3/7 \doteq 0.428\,571$. □

3.G.20. 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, III according to their quality, and a report found out that

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 the class II, the next product is of class II in 60 % of the cases and is of quality I in 20 % of the cases, and

after a product of quality III, the next product is of quality III in 50 % of the cases while in 25 % of the cases it is of quality II.

Compute the probability that the 18-th product is of the quality I, given that the 16-th product is of quality III.

Solution. First we solve the problem without using a Markov chain. Since 16-th product is of the class III, the event in question is satisfied by the cases

- 17-th product is of the class I and 18-th product is of class I;
- 17-th product is of the class II and 18-th product is of class I;
- 17-th product is of the class III and 18-th product is of class I,

with probabilities respectively

- $0.25 \cdot 0.8 = 0.2$;
- $0.25 \cdot 0.2 = 0.05$;
- $0.5 \cdot 0.25 = 0.125$.

Thus the solution is

$$0.375 = 0.2 + 0.05 + 0.125.$$

Now view the problem as a Markov process. From the statement there corresponds the probabilistic matrix

$$\begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix}.$$

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

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

For the next product in order there follows the vector

$$\begin{pmatrix} 0.375 \\ 0.3 \\ 0.325 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix}'.$$

The first component is the desired probability.

Notice that the first method of the solution (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 22-nd or 30-th 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 18-th product belongs to the class II and III. □

3.G.21. Repeated dice casting. Write down the transitional probabilistic matrix T for the Markov chain with states "maximum resulting number after n attempts" with the order of the states $1, \ldots, 6$. Then determine T^n for every $n \in \mathbb{N}$.

Solution. We can immediately write

$$T = \begin{pmatrix} 1/6 & 0 & 0 & 0 & 0 & 0 \\ 1/6 & 2/6 & 0 & 0 & 0 & 0 \\ 1/6 & 1/6 & 3/6 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 4/6 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 5/6 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1 \end{pmatrix}.$$

The first column is determined by the state 1 and probability $1/6$ that it is preserved (that is, the next result is one) and probability $1/6$ for transition into any of the other states $2, \ldots, 6$ (the result on the dice would be $2, \ldots, 6$). The second column is given by the state 2 and probabilities $2/6$ that it is preserved (the result is 1 or 2) and probability for transition $1/6$ for transition into any of the other states $3, \ldots, 6$ (the result would be $3, \ldots, 6$). The last column is derived from the fact that the state 6 is persistent. That is, if 6 has already been cast, no greater result is possible.

For $n \in \mathbb{N}$, we can directly determine T^n:

$$\begin{pmatrix} a^n & 0 & 0 & 0 & 0 & 0 \\ b^n - a^n & b^n & 0 & 0 & 0 & 0 \\ c^n - b^n & c^n - b^n & c^n & 0 & 0 & 0 \\ d^n - c^n & d^n - c^n & d^n - c^n & d^n & 0 & 0 \\ e^n - d^n & e^n - d^n & e^n - d^n & e^n - d^n & e^n & 0 \\ 1 - e^n & 1 - e^n & 1 - e^n & 1 - e^n & 1 - e^n & 1 \end{pmatrix}.$$

where $a = 1/6$, $b = 2/6$, $c = 3/6$, $d = 4/6$, $e = 5/6$.

The numbers in the first column correspond successively to the probabilities that n-times in a row the result is 1, n-times in a row the result is 1 or 2 and there was at least one 2 (therefore we subtract the probability given in the first row), n-times in a row the result is 1, 2 or 3 and at least once the result is 3, up to the last row where there is the probability that at least once during n throws the result is 6 (this can be easily derived from the probability of the complementary event). Similarly, in the fourth column are the non-zero probabilities of the events

"if n-times in a row the result is 1, 2, 3 or 4",

"if n-times in a row the result is 1, 2, 3, 4 or 5 and at least once it is 5"

"if at least once during n attempts the result is 6". Interpretation of the matrix T as the probabilistic transition matrix of a Markov process allows for a quick expression of the powers T^n, $n \in \mathbb{N}$. $\square$

3.G.22. In this problem we deal with a certain property of an animal species which is determined independently of sex but just by a certain gene – a pair of alleles. Every individual gains one allele from each parent, randomly and independently. There are forms of the gene given by various alleles a, A – they form three possible states aa, $aA = Aa$ and AA of the property.

(a) Assume that each individual of a certain population mates only with an individual of another population, where there appears only the property caused by the pair aA. Exactly one of their offspring (a randomly chosen one) will be left on the spot and he will also mate only with an individual of that specific population, and so on. Determine the probabilities of appearance of aa, aA, AA in the considered population after certain time.

(b) Solve the problem given in the case (a), if the other population is composed only of individuals with the pair AA.

(c) Randomly chosen two individuals of opposite sex are bred. From their progeny again randomly choose two of opposite sex and breed them. If this occurs for a long time, compute the probability that both bred individuals have a pair of alleles AA, or aa, when the process of breeding ends.

(d) Solve the problem from case (c) without the condition that the individuals have the same parent. Thus just breed random individuals from a population among them, then breed among their progeny, and so on.

Solution. Case (a). This is a Markov process given by the matrix

$$T = \begin{pmatrix} 1/2 & 1/4 & 0 \\ 1/2 & 1/2 & 1/2 \\ 0 & 1/4 & 1/2 \end{pmatrix}.$$

The order of the states corresponds to the order of the pairs of alleles aa, aA, AA. The numbers in the first column follow from the fact that an offspring of parents with alleles aa and aA has probability $1/2$ for the pair aa and probability $1/2$ for the pair aA. Similarly for the third column. The numbers in the second column follow from the fact that each of the four cases of the pairs of alleles aa, aA, Aa, AA has the same probability for an individual whose both parents have the pair aA.

Note that there is a difference between counting probability — where we must distinguish between aA and Aa (which allele comes from which parent) — and investigating just the properties caused by the pairs aA and Aa which are then the same. For determining the resulting state it thus suffices to find the probabilistic vector associated with the eigenvalue 1 of the matrix T, because the matrix

$$T^2 = \begin{pmatrix} 3/8 & 1/4 & 1/8 \\ 1/2 & 1/2 & 1/2 \\ 1/8 & 1/4 & 3/8 \end{pmatrix}$$

satisfies the condition of the Perron-Frobenius theorem, that is, all of its elements are positive. The probabilistic vector is

$$\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right)^T,$$

which gives the probabilities $1/4$, $1/2$, $1/4$ of the appearance of the combinations aa, aA and AA respectively after a very long (theoretically infinite) time.

Case (b). For the order of the pairs of alleles AA, aA, aa we obtain the probabilistic matrix

$$T = \begin{pmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues are 1, $1/2$ and 0. To these eigenvalues correspond respectively the eigenvectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Therefore

$$T = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

From there for arbitrary $n \in \mathbb{N}$ it follows that

$$T^n = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}^n \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2^{-n} & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since $\lim_{n\to\infty} 2^{-n} = 0$,

$$T^n \approx \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

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

for large n. Thus if individuals of the original population procreate exclusively with the member of the specific population (the one which has only AA), then necessarily after a sufficient amount of breeding there is a total elimination of the pairs aA and aa.

Case (c). There are 6 possible states (in this order)

$$AA, AA; \quad aA, AA; \quad aa, AA;$$

$$aA, aA; \quad aa, aA; \quad aa, aa,$$

while these states are given by the genotypes of the parents. The matrix of the corresponding Markov chain is

$$T = \begin{pmatrix} 1 & 1/4 & 0 & 1/16 & 0 & 0 \\ 0 & 1/2 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1/8 & 0 & 0 \\ 0 & 1/4 & 1 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 & 1/2 & 0 \\ 0 & 0 & 0 & 1/16 & 1/4 & 1 \end{pmatrix}.$$

If we consider for instance the situation (second column), where one of the parents has the pair AA and the other has aA, then each of the four cases (we are talking about the pairs of alleles of two randomly chosen offsprings)

$$AA, AA; \quad AA, aA; \quad aA, AA; \quad aA, aA$$

occurs with the same probability. The probability of staying in the second state is thus $1/2$ and the probability for transition from the second state to the first is $1/4$ and to the fourth state also $1/4$.

Again we determine the powers T^n for large $n \in \mathbb{N}$. Considering the form of the first and of the last column we find that 1 is an eigenvalue of the matrix T with corresponding eigenvectors

$$(1, 0, 0, 0, 0, 0)^T, \quad (0, 0, 0, 0, 0, 1)^T.$$

By considering only a four-dimensional submatrix of the matrix T (omitting the first and sixth row and column) we find the remaining eigenvalues

$$\frac{1}{2}, \quad \frac{1}{4}, \quad \frac{1 - \sqrt{5}}{4}, \quad \frac{1 + \sqrt{5}}{4}.$$

Recalling the solution of the exercise called the sweet-toothed gambler, we do not have to compute T^n. In that exercise we obtained the same eigenvectors corresponding to the eigenvalue 1 and the other eigenvalues also had their absolute value strictly smaller than 1 (the exact values were not used). Thus we obtain an identical conclusion – the process approaches the probabilistic vector

$$(a, 0, 0, 0, 0, 1 - a)^T,$$

where $a \in [0, 1]$ is given by the initial state. Because there is a non zero number only at the first and sixth position of the resulting vector, the states

$$aA, AA; \quad aa, AA; \quad aA, aA; \quad aa, aA$$

disappear after many breedings. Notice that the probability that the process ends with AA, AA equals the relative ratio of the appearance of A in the initial state.

Case (d). Let the values $a, b, c \in [0, 1]$ give, in this order, the relative ratios of the occurrence of alleles AA, aA, aa in the given population. We wish to obtain the expression of relative ratios of the pairs AA, aA, aa in the offspring of the population. If the choice of pairs for breeding is random, then for a suitably large population it can be expected that the relative ratio of breeding of individuals that both have AA is a^2. Simlarly, the relative ratio for the pair aA and AA is $2ab$, and the relative ratio for aA (both of them) is b^2 and so on. The offspring of the parents with pairs AA, AA must inherit AA. The probability that the offspring of the parents with pairs AA, aA has AA is $1/2$ and the probability that the offspring of the parents with pairs aA, aA has AA is $1/4$. There are no other cases for an offspring with the pair AA. If one of the parents has the pair aa, then the offspring cannot have AA). The relative frequency of AA in the progeny is thus

$$a^2 \cdot 1 + 2ab \cdot \frac{1}{2} + b^2 \cdot \frac{1}{4} = a^2 + ab + \frac{b^2}{4}.$$

Similarly we set the relative frequencies of the pairs aA and aa in the progeny:

$$ab + bc + 2ac + \frac{b^2}{2}$$

and

$$c^2 + bc + \frac{b^2}{4}.$$

This process can be viewed as a mapping T that transforms the vector $(a, b, c)^T$. Hence

$$T : \begin{pmatrix} a \\ b \\ c \end{pmatrix} \mapsto \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix}.$$

Note that the domain (and also the codomain) of T are just the vectors

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \text{where} \quad a, b, c \in [0, 1], \ a + b + c = 1.$$

We would like to describe the operation T by multiplying the vector by some constant matrix. But that is clearly not possible since the mapping T is not linear. Thus it is not a Markov process and the determination of what happens after a long time cannot be simplified as in the previous cases. But we can determine what happens if we apply the mapping T twice in succession. For the second step,

$$T : \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix} \mapsto \begin{pmatrix} t_2^1 \\ t_2^2 \\ t_2^3 \end{pmatrix}, \quad \text{where}$$

$$t_2^1 = \left(a^2 + ab + \frac{b^2}{4} \right)^2 +$$
$$\left(a^2 + ab + \frac{b^2}{4} \right) \left(ab + bc + 2ac + \frac{b^2}{2} \right)$$
$$+ \frac{1}{4} \left(ab + bc + 2ac + \frac{b^2}{2} \right)^2,$$

$$t_2^2 = \left(a^2 + ab + \frac{b^2}{4} \right) \left(ab + bc + 2ac + \frac{b^2}{2} \right) +$$
$$+ \left(ab + bc + 2ac + \frac{b^2}{2} \right) \left(c^2 + bc + \frac{b^2}{4} \right) +$$
$$+ 2 \left(a^2 + ab + \frac{b^2}{4} \right) \left(c^2 + bc + \frac{b^2}{4} \right) +$$
$$+ \frac{1}{2} \left(ab + bc + 2ac + \frac{b^2}{2} \right)^2,$$

$$t_2^3 = \left(c^2 + bc + \frac{b^2}{4}\right)^2 +$$

$$\left(ab + bc + 2ac + \frac{b^2}{2}\right)\left(c^2 + bc + \frac{b^2}{4}\right) +$$

$$\frac{1}{4}\left(ab + bc + 2ac + \frac{b^2}{2}\right)^2.$$

Using $a + b + c = 1$,) it can be shown that

$$t_2^1 = a^2 + ab + \frac{b^2}{4}, \ t_2^2 = ab + bc + 2ac + \frac{b^2}{2}, \ t_2^3 = c^2 + bc + \frac{b^2}{4},$$

that is,

$$T : \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix} \mapsto \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix}.$$

We have obtained a surprising result that further application of the transform T does not change the vector obtained in the first step. This means that the appearance of the considered pairs is, after an arbitrary long time, the same as in the first generation of offspring. For a large population, we have thus shown that the evolution takes place during first generation unless there is a mutation or selection. □

3.G.23. Let there be two boxes, which contain between them n white and n black balls. Each box contains 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 in the year 1769) or analogously, as a model of diffusion of gases.

Let the states $0, 1, \ldots, n$ 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 $j \in \{1, \ldots, n\}$ 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 $j \in \{0, \ldots, n-1\}$ 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 $j \in \{1, \ldots, n-1\}$, if from both boxes balls of the same colour are drawn, which has the same probability

$$\frac{j}{n} \cdot \frac{n - j}{n} + \frac{n - j}{n} \cdot \frac{j}{n} = \frac{2j(n - j)}{n^2}.$$

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 we obtain the matrix $n^2 T$:

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ n^2 & 2 \cdot 1(n-1) & 2^2 & \ddots & 0 & 0 & 0 \\ 0 & (n-1)^2 & 2 \cdot 2(n-2) & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 2 \cdot (n-2)2 & (n-1)^2 & 0 \\ 0 & 0 & 0 & \ddots & 2^2 & 2 \cdot (n-1)1 & n^2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

for the order of the states $0, 1, \ldots, n$.

When using this model 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 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.

If we number the individual balls, we would instead of ball drawing draw some of the numbers $1, 2, \ldots, 2n$ and the ball whose number was draw would move to the other ball. We would obtain a Markov process with states $0, 1, \ldots, 2n$ (the number of balls in the first box), where we are not distinguishing the colour any more. This Markov chain is also very important in physics (P. and T. Ehrenfest introduced it in 1907). It is used as a model for interchange of heat between two isolated bodies.

□

3.G.24. 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 bet always only €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 $1 - p$. They play as long as both have some money. If player A has €x at the start of the game, and player B has €y at the start of the game, determine the probability that player A loses all the money he has.

Solution. This problem is 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

$$
(1) \qquad \frac{1 - \left(\frac{p}{1-p}\right)^{y}}{1 - \left(\frac{p}{1-p}\right)^{x+y}}.
$$

We investigate what this value is for specific choices of p, x, y. If player B wants to be almost sure and requires that the probability that player A loses with him €1 000 000 c is at least 0.999, then it suffices for him to have €346 c if $p = 0.495$ (or €1 727 if $p = 0.499$). Therefore it is possible in big casinos that "passionate" players play almost fair games. □

3.G.25. In a certain company there exist two competing departments. The management has decided that every week they will measure relative (with respect to the number of employees) incomes attained by these two departments. 2 employees will then be moved to the more successful department from the other department. This process will go on for as long as both departments have some employees. You have gained a position in this company and you can choose one of these two departments where you will work. You want to choose the department which will not be cancelled due to the employee movement. What will be your choice, if one of the departments has 40 employees, the other 10 employees and you estimate that the second one will have a greater income than the first one in 54 % of the cases? ○

Solutions of the exercises

3.A.2. The daily diet should contain $3.9\,\text{kg}$ of hay and $4.3\,\text{kg}$ of oat. The costs per foal are then €13.82.

3.B.10.
$$x_n = \frac{1}{\sqrt{21}}\left(\frac{3+\sqrt{21}}{2}\right)^n - \frac{1}{\sqrt{21}}\left(\frac{3-\sqrt{21}}{2}\right)^n.$$

3.B.11. $x_n = 2\sqrt{3}\sin(n\cdot(\pi/6)) - 4\cos(n\cdot(\pi/6))$.

3.B.12. $x_n = -3(-1)^n - 2\cos(n\cdot(2\pi/3)) - 2\sqrt{3}\sin(n\cdot((2\pi/3)))$.

3.B.13. $x_n = (-1)^n(-2n^2 + 8n - 7)$.

3.E.2. yes, no, no, yes

3.F.1.

- The claim is true. ($B := A^T A$, $b_{ij} = $ (i-th row of A^T) $\cdot$ (j-th column of A)$= b_{ji} = $ (j-th row of A^T) $\cdot$ (i-th column of A)=(j-th column of A) $\cdot$ (i-th row of A^T)

- The claim is not true. Consider for instance $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

3.F.3.
$$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & -2 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

3.G.5. The Leslie matrix of the given model is (the mortality of the first group is denoted by a)
$$\begin{pmatrix} 0 & 2 & 2 \\ a & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The stagnation condition corresponds to the fact that the matrix has 1 for the eigenvalue, that is, the polynomial $\lambda^3 - 2a\lambda - 2a$ has 1 as its root, that is, $a = 1/4$.

3.G.10. Again it is a special case of the Ruining of the player. It suffices to reformulate the statement accordingly. For $p = 0.47$, $y = 20$ and $x = 20$ from (1) follows the result
$$0.917 \doteq \frac{1 - \left(\frac{0.47}{1-0.47}\right)^{20}}{1 - \left(\frac{0.47}{1-0.47}\right)^{40}}.$$

3.G.14.
$$\begin{pmatrix} \frac{5}{6} & \frac{1}{5} \\ \frac{1}{6} & \frac{4}{5} \end{pmatrix}.$$
The matrix has the dominant eigenvalue 1, the corresponding eigenvector is $(\frac{6}{5}, 1)$. Because the eigenvalue is dominant, the ratio of the viewers stabilises on $6 : 5$.

3.G.17. As in (3.G.16) the game ends after three bets. Thus all the powers of A, starting with A^3, are identical.
$$A^{100} = A^3 = \begin{pmatrix} 1 & 7/8 & 3/4 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/8 & 1/4 & 1/2 & 1 \end{pmatrix}$$

3.G.25. You can use the result of the exercise called Ruining of the player. According to this exercise the probability that the first department is cancelled exercise equals to
$$\frac{1 - \left(\frac{0.46}{1-0.46}\right)^5}{1 - \left(\frac{0.46}{1-0.46}\right)^{25}} \doteq 0.56.$$

It was enough to substitute i $p = 1 - 0.54$, $y = 10/2$ and $x = 40/2$ into (1). It is thus better to choose the smaller department.

Analytic geometry

position, incidence, projection
– and we return to matrices again...
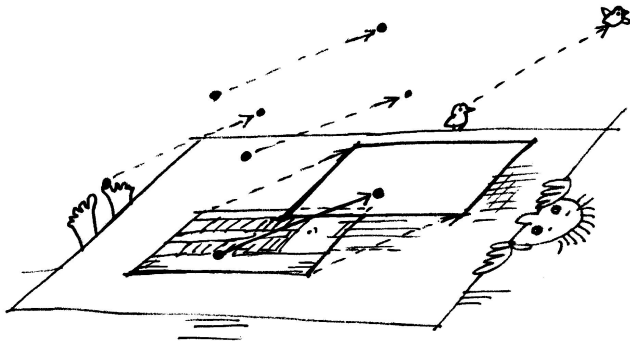
A. Affine geometry

4.A.1. Find a parametric equation for a line in $\mathbb{R}^3$ given by the equations

$$\begin{array}{rcrcrcl} x & - & 2y & + & z & = & 2, \\ 2x & + & y & - & z & = & 5. \end{array}$$

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

$$[2, -1, -2] + t\,(1, 3, 5)\,, \quad t \in \mathbb{R}.$$

$\square$

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.

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.

In the next part, we use linear algebra to study objects defined in a nonlinear way. To do this we nee more from the theory of matrices. The results are important in discussions of the technique of optimization, or of searching for extrema of functions.

At the end of this chapter we show how the projectivization of affine spaces helps us to obtain a 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 first part of the previous 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. A hint as to how to deal with the theory is given already in the discussion of the geometry of the plane, c.f. paragraph 1.5.3 and further on. 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.

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 independent equations. Solve the problem by elimination of parameters. The points $[x, y, z, u] \in \varrho$ satisfy

$$
\begin{array}{rcrcrcr}
x & = & & & t & + & 2s, \\
y & = & 3 & & & - & s, \\
z & = & 2 & + & t & - & 2s, \\
u & = & 5 & & & + & 2s,
\end{array}
$$

where $t, s \in \mathbb{R}$. Write the system as a matrix

$$
\left(\begin{array}{rr|rrrr|r}
1 & 2 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & 0 & 3 \\
1 & -2 & 0 & 0 & -1 & 0 & 2 \\
0 & 2 & 0 & 0 & 0 & -1 & 5
\end{array} \right).
$$

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

$$
\left(\begin{array}{rr|rrrr|r}
1 & 2 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & 0 & 3 \\
0 & 0 & 1 & 4 & -1 & 0 & -10 \\
0 & 0 & 0 & -2 & 0 & -1 & 11
\end{array} \right).
$$

The bottom two rows, both with only zeros to the left of the first vertical line, imply

$$
\begin{array}{rcrcrcrcrcr}
x & + & 4y & - & z & & & - & 10 & = & 0, \\
& & -2y & & & - & u & + & 11 & = & 0.
\end{array}
$$

Note that the original system can be written as

$$
\left(\begin{array}{rrrr|rr|r}
1 & 0 & 0 & 0 & 1 & 2 & 0 \\
0 & 1 & 0 & 0 & 0 & -1 & 3 \\
0 & 0 & 1 & 0 & 1 & -2 & 2 \\
0 & 0 & 0 & 1 & 0 & 2 & 5
\end{array} \right),
$$

where x, y, z, u remains on the left-hand side of the equations. A similar transformation gives

$$
\left(\begin{array}{rrrr|rr|r}
1 & 0 & 0 & 0 & 1 & 2 & 0 \\
0 & 1 & 0 & 0 & 0 & -1 & 3 \\
-1 & -4 & 1 & 0 & 0 & 0 & -10 \\
0 & 2 & 0 & 1 & 0 & 0 & 11
\end{array} \right)
$$

from which

$$
\begin{array}{rcrcrcrcr}
-x & - & 4y & + & z & & & = & -10, \\
& & 2y & & & + & u & = & 11.
\end{array}
$$

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

$$ A + v = (a_1 + v_1, \ldots, a_n + v_n) \in \mathbb{R}^n = \mathcal{A}_n $$

to a point $A = (a_1, \ldots, a_n) \in \mathcal{A}_n$ and a vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n = V$.

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 $\vec{AB}$.

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

$$ \mathcal{P} \times V \to \mathcal{P}, \quad (A, v) \mapsto A + v. $$

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* $\tau_v : \mathcal{A} \to \mathcal{A}$ as the restricted map

$$ \tau_v : \mathcal{P} \simeq \mathcal{P} \times \{v\} \to \mathcal{P}, \quad A \mapsto A + v. $$

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

$$
\begin{array}{rcrcrcl}
2y & & & + & u & = & a, \\
-x & & + z & + & 2u & = & b
\end{array}
$$

yields $a = 11$, $b = 12$. The desired implicit expression is

$$
\begin{array}{rcrcrcl}
2y & & & + & u & = & 11, \\
-x & & + z & + & 2u & = & 12.
\end{array}
$$

Another solution Since

$$
\begin{array}{rclrclrcl}
x & = & & & t & + & 2s, \\
y & = & 3 & & & - & s, \\
z & = & 2 & + & t & - & 2s, \\
u & = & 5 & & & + & 2s,
\end{array}
$$

Eliminate t to get

$$
\begin{array}{rcl}
x - z & = & 2 - 4s, \\
y & = & 3 - s, \\
u & = & 5 + 2s,
\end{array}
$$

Eliminate s to obtain two equations, namely

$$
\begin{array}{rcl}
z - x + 2u & = & 12 \\
u + 2y & = & 11,
\end{array}
$$

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

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 $\mathcal{A}$

(1) $$A - A = 0 \in V$$

(2) $$B - A = -(A - B)$$

(3) $$(C - B) + (B - A) = C - A.$$

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* $(A_0; u_1, \ldots, u_n)$ 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

$$
Z(M) = \langle \{B - A; B, A \in M\} \rangle \subset V
$$

of all vectors generated by the differences of points in M.

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

$$[x, y, z] = [2, 1, 1] + t\,(1, 3, 4) + s\,(2, -3, 2)\,, \quad t \in \mathbb{R},\ s > 0.$$

$\square$

4.A.4. Determine the relative position of the lines

$$p : [1, 0, 3] + t\,(2, -1, -3)\,, \quad t \in \mathbb{R},$$
$$q : [1, 1, 3] + s\,(1, -1, -2)\,, \quad s \in \mathbb{R}.$$

Solution. Search for common points of the given lines (subspaces intersection). We have a system

$$
\begin{array}{ccccccc}
1 & + & 2t & = & 1 & + & s, \\
0 & - & t & = & 1 & - & s, \\
3 & - & 3t & = & 3 & - & 2s.
\end{array}
$$

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

$$
\begin{array}{ccccccc}
4 & + & 2t & = & a & + & s, \\
-4 & + & t & = & 6 & - & 3s, \\
8 & - & 4t & = & -5 & + & 3s
\end{array}
$$

has a solution. Express the system as a matrix (the first column corresponding to t, the second to s), and solve

$$
\begin{pmatrix}
2 & -1 & \vline & a - 4 \\
1 & 3 & \vline & 10 \\
-4 & -3 & \vline & -13
\end{pmatrix}
\sim
\begin{pmatrix}
1 & 3 & \vline & 10 \\
2 & -1 & \vline & a - 4 \\
-4 & -3 & \vline & -13
\end{pmatrix}
$$

$$
\sim
\begin{pmatrix}
1 & 3 & \vline & 10 \\
0 & -7 & \vline & a - 24 \\
0 & 1 & \vline & 3
\end{pmatrix}.
$$

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 $[6, -3, 4]$. $\square$

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 of a subspace

Affine subspaces can be described by their difference spaces after choosing a point $A_0 \in M$ in a generating set M. Indeed, $\langle M \rangle = \{A_0 + v; v \in Z(M) \subset Z(\mathcal{A})\}$. 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. 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 $(u_1, \ldots, u_k)$ be a basis of $Z(\mathcal{Q}) \subset \mathbb{R}^n$. Then the expression of the subspace

$$\mathcal{Q} = \{A + t_1 u_1 + \cdots + t_k u_k; t_1, \ldots, t_k \in \mathbb{R}\}$$

is called the *parametric description* of the subspace $\mathcal{Q}$.

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 $b_i \in \mathbb{R}$, $i = 1, \ldots, n - k$. Suppose $A = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ is a 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

4.A.6. In $\mathbb{R}^3$, determine the relative position of the line p defined implicitly by

$$\begin{array}{rcrcrcr} x & + & y & - & z & = & 4, \\ x & - & 2y & + & z & = & -3 \end{array}$$

and the plane $\varrho : y = 2x - 1$.

Solution. A normal vector to the plane is ϱ 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 ϱ is a linear combination of the p normal vectors. A vector defining the line lies in a subspace of the plane ϱ It remains to discover whether or not they intersect. The system of equations

$$\begin{array}{rcrcrcr} x & + & y & - & z & = & 4, \\ x & - & 2y & + & z & = & -3, \\ 2x & - & y & & & = & 1 \end{array}$$

has infinitely many solutions, because the first two equations add to give the third one. So the line p lies in the plane ϱ. $\square$

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.

4.A.7. Find the intersection of the subspaces Q_1 and Q_2, where

$$Q_1 : [4, -5, 1, -2] + t_1\,(3, 5, 4, 2) + t_2\,(2, 4, 5, 1)$$
$$+ t_3\,(0, 3, 1, 2)\,,$$
$$Q_2 : [4, 4, 4, 4] + s_1\,(0, -6, -2, -4) + s_2\,(-1, -5, -3, -3)\,,$$

$$\text{for } t_1, t_2, t_3, s_1, s_2 \in \mathbb{R}.$$

Solution. The point $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^4$ lies in Q_1 if and only if

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ -5 \\ 1 \\ -2 \end{bmatrix} + t_1 \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \end{pmatrix} + t_2 \begin{pmatrix} 2 \\ 4 \\ 5 \\ 1 \end{pmatrix} + t_3 \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix}$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix}$$

k. The set of all solutions of the given system is of the form $\{B; B = A + (y_1, \ldots, y_n)^T, y = (y_1 \ldots, y_n)^T \in U\} \subset \mathbb{R}^n$, 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 $\mathcal{Q} \subset \mathcal{A}_n$. 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 $\mathcal{Q} = B + Z(\mathcal{Q})$, it is necessary to describe the difference space of the subspace $\mathcal{Q}$ 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 $Z(\mathcal{Q})$. In these coordinates, the vectors $v \in Z(\mathcal{Q})$ are given by equations

$$\alpha_j(v) = 0, \quad j = k + 1, \ldots, n.$$

The α_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 $Z(\mathcal{Q})$ 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 $X \in \mathcal{A}$

$$\begin{aligned} X &= B_0 + x_1' v_1 + \cdots + x_n' v_n \\ &= B_0 + (A_0 - B_0) + x_1 u_1 + \cdots + x_n u_n. \end{aligned}$$

Let $y = (y_1, \ldots, y_n)^T$ 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

$$x_1' = y_1 + a_{11}x_1 + \cdots + a_{1n}x_n$$
$$\vdots$$
$$x_n' = y_n + a_{n1}x_1 + \cdots + a_{nn}x_n.$$

In matrix notation

$$x' = y + M \cdot x.$$

for some $s_1, s_2 \in \mathbb{R}$. Hence if X lies in $Q_1 \cap Q_2$ then the equation

$$t_1 \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \end{pmatrix} + t_2 \begin{pmatrix} 2 \\ 4 \\ 5 \\ 1 \end{pmatrix} + t_3 \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix}$$
$$= \begin{pmatrix} 4-4 \\ 4+5 \\ 4-1 \\ 4+2 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix}.$$

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

$$\begin{pmatrix} 3 & 2 & 0 & 0 & 1 & 0 \\ 5 & 4 & 3 & 6 & 5 & 9 \\ 4 & 5 & 1 & 2 & 3 & 3 \\ 2 & 1 & 2 & 4 & 3 & 6 \end{pmatrix} \sim \begin{pmatrix} 3 & 2 & 0 & 0 & 1 & 0 \\ 0 & 2 & 9 & 18 & 10 & 27 \\ 0 & 7 & 3 & 6 & 5 & 9 \\ 0 & -1 & 6 & 12 & 7 & 18 \end{pmatrix}$$

$$\sim \cdots \sim \left(\begin{array}{ccccc|c} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right).$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + t \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix}.$$

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 $Q_1 \cap Q_2$. $\square$

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.

For example, the influence of such a change of basis on the coordinates of subsets is described by systems of linear equations. Suppose the system in coordinates $(A_0; \underline{u})$ has the form

$$S \cdot x = b$$

where S is the matrix of the system. Then

$$S \cdot x = S \cdot M^{-1} \cdot (y + M \cdot x) - S \cdot M^{-1} \cdot y = b.$$

Thus in the new coordinates $(B_0; \underline{v})$ considered above, the system has the form

$$(S \cdot M^{-1}) \cdot x' = b' = b + (S \cdot M^{-1}) \cdot y.$$

Therefore, if a subset is described by a system of linear equations in one affine frame, then it is so described in the all other affine frames. This completes the proof of the previous proposition.

4.1.6. Examples of affine subspaces. (1) The one-dimensional (standard) affine space is the subset 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 $(a_1, \ldots, a_n)$ 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 $[x_1, \ldots, x_n] \in \mathcal{A}_n$. This is an affine subspace of dimension $n - 1$. We say that the subspace is of codimension 1, called a *hyperplane in A_n*.

4.1.7. Affine combinations of points. We introduce an analogue of the linear combination of vectors. Let $A_0, \ldots, A_k$ be points in the affine space $\mathcal{A}$. Their affine hull $\langle \{A_0 \ldots, A_k\} \rangle$ can be written as

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

$$[2, 1, 0, 1] = t_1[0, 2, 1, 1] + t_2[1, 0, 1, 1] + t_3[3, -2, -1, 1]$$
$$+ t_4[-1, 0, 1, 1]$$

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 $A = [4, 0, 2]$, $B = [-2, -3, 1]$, $C = [1, -1, -3]$, $D = [2, 4, -2]$.

 a) Determine its volume.

 b) 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. $\square$

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 $\{[0, 0, 0], (1, 0, 0), (0, 1, 0), (0, 0, 1)\}$.

$$\{A_0 + t_1(A_1 - A_0) + \cdots + t_k(A_k - A_0); t_1, \ldots, t_k \in \mathbb{R}\}.$$

In any affine coordinates the same set can be written as

$$\langle A_0, \ldots, A_k \rangle = \{t_0 A_0 + t_1 A_1 + \cdots + t_k A_k; t_i \in \mathbb{R}, \sum_{i=0}^{k} t_i = 1\}.$$

AFFINE COMBINATIONS OF POINTS

In general, by the formula $t_0 A_0 + t_1 A_1 + \cdots + t_k A_k$ with coefficients satisfying $\sum_{i=0}^{k} t_i = 1$ is meant the points $A_0 + \sum_{i=1}^{k} t_i (A_i - A_0)$. They are called the *affine combinations of points*.

The points $A_0 \ldots, A_k$ 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 $A_0 \ldots, A_k$ 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 $A_0, \ldots, A_k$ be $k + 1$ points in general position in an affine space $\mathcal{A}$. The set $\Delta = \Delta(A_0, \ldots, A_k)$ is defined as the set of all affine combinations of points A_i with nonnegative coefficients only. This is

$$\Delta = \{t_0 A_0 + t_1 A_1 + \cdots + t_k A_k; t_i \in [0, 1] \subset \mathbb{R}, \sum_{i=0}^{k} t_i = 1\},$$

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 $t_i = 0, i = 0, \ldots, k$. 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.

Solution. The coordinates $[2, 2, 3]$ in the given basis are by the definition

$$[1, 2, 3] + 2 \cdot (1, 1, 1) + 2 \cdot (1, -1, 2) + 3 \cdot (2, 1, 1)$$
$$= [11, 5, 12]$$

coordinates for X in the standard basis. □

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

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1} \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \end{pmatrix} \cdot \right)$$
$$= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Hence the desired mapping is

$$f(y_1, y_2)$$
$$= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left[\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \left(\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right) \right.$$
$$\left. + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] + \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$
$$= \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

□

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

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 also the entire simplex defined by these points

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 *subspaces*

$$\alpha = \{P + t_1 \cdot v_1 + \cdots + t_k \cdot v_k; \ t_1, \ldots, t_k \in \mathbb{R}, t_k \geq 0\},$$

(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*

$$\mathcal{K}(M) = \{t_1 A_1 + \cdots + t_s A_s; \ \sum_{i=1}^{s} t_i = 1, \ t_i \geq 0, A_i \in M\}$$

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, $i = 1, .., s_1$, t'_j, $j = 1, \ldots, s_2$ 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:

$$\varepsilon(t_1 A_1 + \cdots + t_s A_s) + (1 - \varepsilon)(t'_1 A_1 + \cdots + t'_s A_s), \ 0 \leq \varepsilon \leq 1.$$

Obviously any point of this line segment lies in S.

It remains to show that the convex hull of the points $A_1, \ldots, A_s$ 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 $A_1, \ldots, A_{s-1}$ 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 $A = t_1 A_1 + \cdots + t_s A_s \in S, t_s < 1$, and affine combinations

$$\varepsilon(t_1 A_1 + \cdots + t_{s-1} A_{s-1}) + (1 - \varepsilon(1 - t_s)) A_s, \ 0 \leq \varepsilon \leq \tfrac{1}{1 - t_s}.$$

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 $A_1, \ldots, A_{s-1}$). 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 $A_1, \ldots, A_s$. □

The convex hulls of finite sets are called *convex polyhedrons*.

237

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 $(-1,2,2)$. $\quad\square$

4.A.15. Find a transversal of the lines (that is, a line passing through both given lines)

$p : [1,1,1] + t(2,1,0), \quad q : [2,2,0] + t(1,1,1),$

so that $[1,0,0]$ lies on the transversal.

Solution. The transversal lies in the plane ρ 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 ρ 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 $P = [7/3, 5/3, 1]$. $\quad\square$

4.A.16. Find the common perpendicular between the two skew lines

$$\begin{aligned} p : & \quad [3,0,3] + (0,1,2)t, \quad t \in \mathbb{R} \\ q : & \quad [0,-1,-2] + (1,2,3)s \quad s \in \mathbb{R} \end{aligned}$$

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 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

$$\mathcal{P}_k(A; u_1, \dots, u_k) = \{A + c_1 u_1 + \dots + c_k u_k; \ 0 \le c_i \le 1\}.$$

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 exercises. (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). To find a plane in R_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 a subspace $\mathcal{Q}_i$ and by the sum of all difference spaces. For instance,

$$\mathcal{Q} = A_1 + (Z(\{A_1, \dots, A_k\}) + Z(\mathcal{Q}_1) + \dots + Z(\mathcal{Q}_s)).$$

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 then two, we must search for the intersection step by step.

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 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. $\qquad\square$

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:

$$
\rho(p, q) = \frac{|(-1, -6, 1) \cdot (4, -1, 2)|}{\|(4, -1, 2)\|} = \frac{4}{\sqrt{21}}.
$$

$\qquad\square$

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 $C = [-5, -13, -2]$.

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 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 $\mathcal{Q}$ 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 $q \subset \mathcal{Q}$. Otherwise we consider the intersection $\mathcal{Q}$ 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. Remarks 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

$$
a_1 x_1 + \cdots + a_n x_n \leq b
$$

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)). Suppose we choose the parametric description of the bounding hyperplane

$$
\{P + t_1 v_1 + \cdots + t_{n-1} v_{n-1}\}
$$

with vectors $v_1, \ldots, v_{n-1}$ from the difference space. By completing these vectors by v to a basis of the whole $\mathbb{R}^n$, the value of

$$
a_1 x_1 + \cdots + a_n x_n - b
$$

on the linear combination $t_1 v_1 + \cdots + t_{n-1} v_{n-1} + t_n v$ must be positive for all vectors with either a positive or a negative t_n.

At the same time, 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

Solution. First, express the line p parametrically. Solve the system

$$
\begin{array}{rcrcrcr}
x & + & 2y & + & z & = & 1, \\
3x & - & y & + & 4z & = & 29.
\end{array}
$$

Rewrite the system as an augmented matrix and perform row operations

$$
\begin{pmatrix} 1 & 2 & 1 & \bigm| & 1 \\ 3 & -1 & 4 & \bigm| & 29 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 1 & \bigm| & 1 \\ 0 & -7 & 1 & \bigm| & 26 \end{pmatrix}
$$
$$
\sim \begin{pmatrix} 1 & 0 & 9/7 & \bigm| & 59/7 \\ 0 & 1 & -1/7 & \bigm| & -26/7 \end{pmatrix}.
$$

The line p is thus described by

$$
p : \left[\frac{59}{7}, -\frac{26}{7}, 0 \right] + t \left(-\frac{9}{7}, \frac{1}{7}, 1 \right), \quad t \in \mathbb{R}.
$$

It is convenient to avoid the fractions, by introducing the substitution $t = 7s + 26$. p is thus described by

$$
p : [-25, 0, 26] + s \left(-9, 1, 7 \right), \quad s \in \mathbb{R}.
$$

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

$$
\sqrt{(-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2}
$$
$$
= \sqrt{(-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2},
$$

or rather

$$
(-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2
$$
$$
= (-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2.
$$

which has the unique solution $s = -3$. Therefore

$$
A = [-25, 0, 26] - 3 \left(-9, 1, 7 \right) = [2, -3, 5].
$$

$\square$

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 $[2, 1, 2]$?

$$
\begin{array}{rcl}
p & : & [-1, 4, 1] + t(-1, 2, 0), \\
q & : & [4, 4, -1] + s(1, 2, -4)?
\end{array}
$$

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. $\square$

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. An example of a practical algorithm for finding the corresponding vertices of a convex polyhedron is given in the chapter about discrete mathematics.

4.1.12. Affine maps. A map $f : \mathcal{A} \to \mathcal{B}$ between affine spaces is called an *affine map* if there exists a linear map $\varphi : Z(\mathcal{A}) \to Z(\mathcal{B})$ 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 φ 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 $t_0 A_0 + \cdots + t_s A_s \in \mathcal{A}$ we obtain

$$
\begin{aligned}
f(t_0 A_0 + \cdots + t_s A_s) & = \\
& = f(A_0 + t_1(A_1 - A_0) + \cdots + t_s(A_s - A_0)) \\
& = f(A_0) + t_1 \varphi(A_1 - A_0) + \cdots + t_s \varphi(A_s - A_0) \\
& = t_0 f(A_0) + t_1 f(A_1) + \cdots + t_s f(A_s).
\end{aligned}
$$

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 $t_i = 1$, $i = 1, \ldots, s$, we define the map φ 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 φ. 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 φ 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 $A_0, \ldots, A_{k+1}$ can be expressed as

$$
r(t_0 A_0 + \cdots + t_k A_k) + s A_{k+1},
$$

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.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

$$U : 4x_1 - 2x_2 - 3x_3 - 2x_4 + 12 = 0,$$

$$2x_1 - x_2 - 2x_3 - 2x_4 + 9 = 0.$$

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

$$((A - B) - P_{A-B}) \perp (4, -2, -3, -2),$$

$$((A - B) - P_{A-B}) \perp (2, -1, -2, -2).$$

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

$$(2, -8, 1, 1)\cdot(4, -2, -3, -2)$$

$$-a(4, -2, -3, -2)\cdot(4, -2, -3, -2)$$

$$-b(2, -1, -2, -2)\cdot(4, -2, -3, -2) = 0,$$

$$((2, -8, 1, 1)\cdot(2, -1, -2, -2))$$

$$-a(4, -2, -3, -2)\cdot(2, -1, -2, -2)$$

$$-b(2, -1, -2, -2)\cdot(2, -1, -2, -2 = 0.$$

If we compute these dot products, we obtain the system

$$\begin{array}{ccccccc}
19 & - & 33a & - & 20b & = & 0, \\
8 & - & 20a & - & 13b & = & 0,
\end{array}$$

4.1.13. 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

$$C = A + s(B - A) = B + r(A - B),$$

the ratio λ 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.*

4.1.14. Changes of coordinates. Under the choice of an affine coordinate system $(A_0, \underline{u})$ on $\mathcal{A}$ and a system $(B_0, \underline{v})$ on $\mathcal{B}$, we obtain the coordinate expression of the affine map $f : \mathcal{A} \to \mathcal{B}$. It is sufficient to express the image $f(A_0)$ of the origin of the coordinate system on $\mathcal{A}$ in the coordinate system on $\mathcal{B}$. In other words, the vector $f(A_0) - B_0$ with the basis $\underline{v}$ is expressed as a column of coordinates y_0. Everything else is then given by multiplying by the matrix of the map φ 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 φ.

As in the case of linear maps, the transformation of affine coordinates corresponds to the expression of the identity map in the chosen affine frames. The change of coordinate expression of an affine map caused by a change of the basis is computed by multiplying and adding matrices and vectors.

Let

$$x = w + M \cdot x',$$

describe a change of basis on the domain by a translation w and a matrix M.

Let

$$y' = z + N \cdot y$$

describe a change of basis on the range space by a translation z and a matrix N.

Then

$$y' = z + N \cdot y = z + N \cdot (y_0 + Y \cdot x)$$

$$= (z + N \cdot y_0 + N \cdot Y \cdot w) + (N \cdot Y \cdot M) \cdot x'.$$

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 $N \cdot Y \cdot M$.

with the only solution $a = 3$, $b = -4$. Hence

$$P_{A-B} = 3\,(4, -2, -3, -2) - 4\,(2, -1, -2, -2)$$
$$= (4, -2, -1, 2),$$

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

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & x_3 & = & 0, \\ & & x_2 & - & x_3 & = & 0, \\ x_1 & + & x_2 & + & 2x_3 & = & 6, \\ x_1 & & & + & 3x_3 & = & 0, \end{array}$$

by the least squares method. (Note that the system does not have a solution – the distance would be 0 otherwise.) Multiply $A \cdot x = b$ 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

$$\left(\begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 3 & 6 & 3 & 6 \\ 6 & 3 & 15 & 12 \end{array} \right) \sim \left(\begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 0 & 3 & -3 & 0 \\ 0 & -3 & 3 & 0 \end{array} \right)$$

$$\sim \left(\begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right).$$

Continue with backward elimination

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0 & 3 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right).$$

4.1.15. 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.

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:

4.1.16. 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 $A = A_0 + a_1 e_1 + \cdots + a_n e_n$, $B = A_0 + b_1 e_1 + \cdots + b_n e_n$ 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 $Z(\mathcal{Q})^{\perp}$ for arbitrary points $B \in \mathcal{Q}$ and $A \in \mathcal{R}$.*

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

$$v = ||\,[2, 0, 2, 2] - [0, 0, 6, 0]\,|| = \sqrt{2^2 + 0 + (-4)^2 + 2^2}$$
$$= 2\sqrt{6}.$$

$\square$

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 ± 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

$$\begin{vmatrix} 4 & -2 \\ 1 & 3 \end{vmatrix} = 14.$$

$\square$

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 $t = 1/3$.

$\square$

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 is because any potential change of B is apparent by adding a vector from $Z(\mathcal{Q})$.

Choose $P = A + (-u_2) = B + u_1 \in \mathcal{Q}$. 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 $\|u_2\|$. $\square$

We consider more elementary problems in affine geometry.

4.1.17. Examples of standard problems. (1) *To find the distance from the point $A \in \mathcal{E}_n$ to the subspace $\mathcal{Q} \subset \mathcal{E}_n$:*

A method of solving such a problem is given in proposition 4.1.16.

(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. 2.3.21). 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 4.1.16.

(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 4.1.16 (the last item). Let η 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

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 φ 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 $\varphi = \pi/3$. $\qquad\square$

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 φ is then

$$\cos(\varphi) = \left| \frac{(1/2,0,-1)\cdot(0,1,1)}{\|(1/2,0,-1)\|\,\|(0,1,1)\|} \right| = \frac{\sqrt{2}}{\sqrt{5}}$$

. $\qquad\square$

4.B.10. Compute the angle between the line p given by the implicit equations

$$\begin{array}{rcrcrcl} x &+& 3y &+& z &=& 0, \\ -x &-& y &+& z &=& 0 \end{array}$$

and plane the $\varrho : x + y + 2z + 1 = 0$.

Solution. The normal vector of the plane ϱ is $(1,1,2)$. Copy the first equation of the line p. Sum both of them, to obtain

$$\begin{array}{rcrcrcl} x &+& 3y &+& z &=& 0, \\ && 2y &+& 2z &=& 0. \end{array}$$

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 φ between the vectors $(1,1,2)$, $(2,-1,1)$,

$$\cos\varphi = \frac{2-1+2}{\sqrt{6}\cdot\sqrt{6}} = \frac{1}{2}.$$

Hence $\varphi = 60°$. However, this is the angle between the direction vector of p and the normal vector ϱ. The desired angle is the complement of this angle, so the solution is $30° = 90° - 60°$. $\qquad\square$

not parallel, η 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.

4.1.18. 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 2.3.21.

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

$$\|u-v\|^2 = \|u\|^2 + \|v\|^2 - 2(u\cdot v)$$
$$= \|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\cos\varphi(u,v).$$

This is the well known *cosine rule* from plane geometry.

The following relation holds for each orthonormal basis $\underline{e}$ of the difference space V and a non–zero vector $u \in V$

$$\|u\|^2 = \sum_i |u\cdot e_i|^2.$$

By dividing this equation by the number $\|u\|^2$,

$$1 = \sum_i (\cos\varphi(u,e_i))^2,$$

which is the law of direction cosines $\varphi(u,e_i)$ of the vector u.

Now we can choose definitions for angles between general subspaces in a Euclidean vector space from the definitions of angles between vectors. Concurrently 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:

4.B.11. In the real plane, find a line which passes through the point $[-3, 0]$, so that $60°$ 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 $-30°$ from the positive x axis. Thus the required line is either at angle $-90°$ or angle $30°$ 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

$$ax + by + c = 0. \quad \text{Choose parameters so that} \quad a^2 + b^2 = 1.$$

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 $60°$ then gives

$$\frac{1}{2} = \cos 60° = \frac{|\sqrt{3}a + 3b|}{\sqrt{12}}, \quad \text{tj.} \quad \sqrt{3} = |\sqrt{3}a + 3b|.$$

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

$$0 = 2b^2 + 2\sqrt{3}ab, \quad \text{tj.} \quad 0 = b\left(b + \sqrt{3}a\right).$$

Together (remember that $c = 3a$ and $a^2 + b^2 = 1$)

$$a = \pm 1, \quad b = 0, \quad c = \pm 3; \qquad a = \pm\frac{1}{2}, \quad b = \mp\frac{\sqrt{3}}{2}, \quad c = \pm\frac{3}{2}.$$

We can easily check that lines determined by those coefficients

$$x + 3 = 0, \qquad \frac{1}{2}x - \frac{\sqrt{3}}{2}y + \frac{3}{2} = 0$$

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 $60°$, and further, that they contain the line $p : [1, 0, 0] + t(1, 1, 0)$. $\bigcirc$

4.B.13. Determine the angle between the planes

$$\sigma : \quad [1, 0, 2] + (1, -1, 1)t + (0, 1, -2)s$$
$$\rho : \quad [3, 3, 3] + (1, -2, 0)t + (0, 1, 1)s$$

Solution. The line of intersection between the planes has direction vector $(1, -1, 1)$. The plane orthogonal to this vector

4.1.19. 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

$$\cos \alpha = \frac{|u.v|}{\|u\|\|v\|}.$$

(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:

4.1.20. 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 φ 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,

$$\frac{|u \cdot v|}{\|u\|\|v\|} = \frac{|u \cdot (v_1 + v_2)|}{\|u\|\|v\|} = \frac{|u \cdot v_1|}{\|u\|\|v\|}$$

$$\leq \frac{\|u\|\|v_1\|}{\|u\|\|v\|} = \frac{\|v_1\|}{\|v\|} = \frac{\|v_1\|^2}{\|v\|\|v_1\|} = \frac{|v_1 \cdot v|}{\|v\|\|v_1\|}.$$

has intersection with the given planes generated by the vectors vektory $(1, 0, -1)$ and $(0, 1, 1)$. The angle between these one-dimensional subspaces is $60°$. □

4.B.14. A cube $ABCDA'B'C'D'$ is given in standard notation. That is, $ABCD$ and $A'B'C'D'$ are faces and AA', BB' are edges. Compute the angle φ 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 $\varphi = 60°$.

□

For further exercises on angles, see .

4.B.15. Prove that for every $n \in \mathbb{N}$ and for all positive $x_1, x_2, \ldots, x_n \in \mathbb{R}$

$$n^2 \le \left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n} \right) \cdot (x_1 + x_2 + \cdots + x_n).$$

For what arguments does equality hold?

Solution. It is sufficient to consider the Cauchy inequality

$$| u \cdot v | \le \| u \| \| v \|$$

in Euclidean space $\mathbb{R}^n$ for the vectors

$$u = \left(\frac{1}{\sqrt{x_1}}, \ldots, \frac{1}{\sqrt{x_n}} \right), v = (\sqrt{x_1}, \ldots, \sqrt{x_n}).$$

We get

$$(1) \quad n \le \sqrt{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \cdot \sqrt{x_1 + x_2 + \cdots + x_n}.$$

We obtain the desired inequality squaring (1). The Cauchy inequality attains equality when vector u is a multiple of v, that is, when $x_1 = x_2 = \cdots = x_n$. □

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 $\mathcal{P}_3(0; \underline{u}, \underline{v}, \underline{t})$

for all vectors $u \in U$. This implies that

$$\cos \varphi(\langle v \rangle, \langle u \rangle) \le \cos \varphi(\langle v \rangle, \langle v_1 \rangle) = \frac{\|v_1\|}{\|v\|}.$$

Thus the computed vector v_1 represents the largest possible value of the cosine of angles between all choices of vectors in U. The cosine 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. □

4.1.21. 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 $U_1 = \langle e_1, \ldots, e_k \rangle$, $U_2 = \langle e'_1, \ldots, e'_l \rangle$.

Consider the orthogonal projection φ of the space V on U_2. Its restriction on U_1 will be denoted by $\varphi : U_1 \to U_2$ as before. Similarly, let $\psi : U_2 \to U_1$ be the map which has arisen from the orthogonal projection on U_1. In the bases $(e_1, \ldots, e_k)$ and $(e'_1, \ldots, e'_l)$, these maps have matrices

$$A = \begin{pmatrix} e_1 \cdot e'_1 & \ldots & e_k \cdot e'_1 \\ \vdots & & \vdots \\ e_1 \cdot e'_l & \ldots & e_k \cdot e'_l \end{pmatrix}, B = \begin{pmatrix} e'_1 \cdot e_1 & \ldots & e'_l \cdot e_1 \\ \vdots & & \vdots \\ e'_1 \cdot e_k & \ldots & e'_l \cdot e_k \end{pmatrix}.$$

$e_i \cdot e'_j = e'_j \cdot e_i$ holds for all indices i, j. Consequently $B = A^T$.

The composition of maps $\psi \circ \varphi : U_1 \to U_1$ has therefore a symmetric positive semidefinite matrix $A^T A$, and ψ is adjoint to φ. 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 λ 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 λ. Consider all eigenvalues $\lambda_1, \ldots, \lambda_k$ (including multiplicities), and let $\underline{u} = (u_1, \ldots, u_n)$ be the corresponding orthonormal basis of U_1 containing the eigenvectors. $\|u\| = 1$, and assume that $\lambda = \lambda_1, u = u_1$.

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

is the absolute value of determinant

$$
\begin{vmatrix} u_1 & v_1 & t_1 \\ u_2 & v_2 & t_2 \\ u_3 & v_3 & t_3 \end{vmatrix} = \begin{vmatrix} t_1 & t_2 & t_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \underline{t} \cdot (\underline{u} \times \underline{v}) \le \|\underline{t}\| \|\underline{u} \times \underline{v}\|
$$

$$
= \|\underline{u} \times \underline{v}\|.
$$

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

$$
t = \pm \frac{(\underline{u} \times \underline{v})}{\|(\underline{u} \times \underline{v})\|}.
$$

$\square$

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

$$
\varrho_1 : [7, 2, 7, -1, 1] + t_1 (1, 0, -1, 0, 0) + s_1 (0, 1, 0, 0, -1),
$$

$$
\varrho_2 : [2, 4, 7, -4, 2] + t_2 (1, 1, 1, 0, 1) + s_2 (0, -2, 0, 0, 3),
$$

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.

$$
\begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & -2 & 0 & 0 & 3 \end{pmatrix} \sim \cdots \sim \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.
$$

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. $\square$

cannot be greater. By the previous lemma, it is sufficient to discuss the angle between u and $\varphi(u) \in U_2$. Choose $v \in U_1$, $v = a_1 u_1 + \cdots + a_k u_k$, $\sum_{i=1}^{k} a_i^2 = \|v\|^2 = 1$. Then

$$
\|\varphi(v)\|^2 = \varphi(v) \cdot \varphi(v) = (\psi \circ \varphi(v)) \cdot v
$$

$$
\le \|\psi \circ \varphi(v)\| \|v\| = \|\psi \circ \varphi(v)\|.
$$

Moreover, the previous lemma gives a formula for computing the angle α between the vector v and the subspace U_2

$$
\cos \alpha = \frac{\|\varphi(v)\|}{\|v\|} = \|\varphi(v)\|.
$$

Since λ_1 is the largest eigenvalue, and the sum of squares of coordinates a_i^2 is one,

$$
(\cos \alpha)^2 = \|\varphi(v)\|^2 \le \|\psi \circ \varphi(v)\| = \sqrt{\sum_{i=1}^{k} (\lambda_i a_i)^2} =
$$

$$
= \sqrt{\lambda_1^2 + \sum_{i=1}^{k} a_i^2 (\lambda_i^2 - \lambda_1^2)} \le \sqrt{\lambda_1^2}.
$$

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. $\square$

4.1.22. 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 $\mathbb{R}^n$.

Let $u_1, \ldots, u_k$ 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 $\mathcal{P}_k(A; u_1, \ldots, u_k) \subset \mathcal{E}_n$ is given by

$$
\mathcal{P}_k(A; u_1, \ldots, u_k) = \{A + c_1 u_1 + \cdots + c_k u_k; 0 \le c_i \le 1\}.
$$

4.B.19. In Euclidean space $\mathbb{R}^5$ determine the distance of planes

$$\sigma_1 : [0, 1, 2, 0, 0] + p_1\,(2, 1, 0, 0, 1) + q_1\,(-2, 0, 1, 1, 0)\,,$$

$$\sigma_2 : [3, -1, 7, 7, 3] + p_2\,(2, 2, 4, 0, 3) + q_2\,(2, 0, 0, -2, -1)\,,$$

where $p_1, q_1, p_2, q_2 \in \mathbb{R}$.

Solution. The sum of the directions σ_1, σ_2 is generated by the direction vectors. Denote them by

$$u_1 = (2, 1, 0, 0, 1)\,, \quad u_2 = (-2, 0, 1, 1, 0)\,,$$
$$v_1 = (2, 2, 4, 0, 3)\,, \quad v_2 = (2, 0, 0, -2, -1)\,.$$

Find points $X_1 \in \sigma_1$, $X_2 \in \sigma_2$, that equal the distance between σ_1 and σ_2. This requires

$$X_1 - X_2 = [0, 1, 2, 0, 0] - [3, -1, 7, 7, 3]$$
$$+ p_1 u_1 + q_1 u_2 - p_2 v_1 - q_2 v_2$$

and

$$\langle\, X_1 - X_2, u_1 \,\rangle = 0, \quad \langle\, X_1 - X_2, u_2 \,\rangle = 0,$$
$$\langle\, X_1 - X_2, v_1 \,\rangle = 0, \quad \langle\, X_1 - X_2, v_2 \,\rangle = 0.$$

Hence

$$\langle\, (-3, 2, -5, -7, -3), u_1 \,\rangle + p_1 \langle\, u_1, u_1 \,\rangle + q_1 \langle\, u_2, u_1 \,\rangle$$
$$- p_2 \langle\, v_1, u_1 \,\rangle - q_2 \langle\, v_2, u_1 \,\rangle = 0,$$
$$\langle\, (-3, 2, -5, -7, -3), u_2 \,\rangle + p_1 \langle\, u_1, u_2 \,\rangle + q_1 \langle\, u_2, u_2 \,\rangle$$
$$- p_2 \langle\, v_1, u_2 \,\rangle - q_2 \langle\, v_2, u_2 \,\rangle = 0,$$
$$\langle\, (-3, 2, -5, -7, -3), v_1 \,\rangle + p_1 \langle\, u_1, v_1 \,\rangle + q_1 \langle\, u_2, v_1 \,\rangle$$
$$- p_2 \langle\, v_1, v_1 \,\rangle - q_2 \langle\, v_2, v_1 \,\rangle = 0,$$
$$\langle\, (-3, 2, -5, -7, -3), v_2 \,\rangle + p_1 \langle\, u_1, v_2 \,\rangle + q_1 \langle\, u_2, v_2 \,\rangle$$
$$- p_2 \langle\, v_1, v_2 \,\rangle - q_2 \langle\, v_2, v_2 \,\rangle = 0.$$

By computing the dot products, we obtain the linear equation system

$$
\begin{array}{rcrcrcrcr}
6p_1 & - & 4q_1 & - & 9p_2 & - & 3q_2 & = & 7, \\
-4p_1 & + & 6q_1 & & & + & 6q_2 & = & 6, \\
9p_1 & & & - & 33p_2 & - & q_2 & = & 31, \\
3p_1 & - & 6q_1 & - & p_2 & - & 9q_2 & = & -11.
\end{array}
$$

Solve it by forming a matrix and performing elementary row operations.

$$
\left(
\begin{array}{rrrr|r}
6 & -4 & -9 & -3 & 7 \\
-4 & 6 & 0 & 6 & 6 \\
9 & 0 & -33 & -1 & 31 \\
3 & -6 & -1 & -9 & -11
\end{array}
\right) \sim \cdots
$$

If the vectors $u_1, \ldots, u_k$ are linearly dependent, we have a k–dimensional parallelepiped $\mathcal{P}_k(A; u_1 \ldots, u_k) \subset \mathcal{E}_n$. For given vectors $u_1, \ldots, u_k$ there are also parallelepipeds of lower dimension

$$\mathcal{P}_1(A; u_1), \ldots, \mathcal{P}_k(A; u_1, \ldots, u_k)$$

in Euclidean subspaces $A + \langle u_1 \rangle, \ldots, A + \langle u_1, \ldots, u_k \rangle$ at our disposal.

If $u_1, \ldots, u_k$ are linearly independent, the volume is given by

$$\mathrm{Vol}\, P_k = 0.$$

Otherwise consider it as in the case of the Gramm–Schmidt orthogonalization

$$\langle u_1, \ldots, u_k \rangle =$$
$$= \langle u_1, \ldots, u_{k-1} \rangle \oplus \langle u_1, \ldots, u_{k-1} \rangle^\perp \cap \langle u_1, \ldots, u_k \rangle.$$

In this decomposition, u_k is uniquely expressed as

$$u_k = u_k' + e_k$$

where $e_k \perp \langle u_1, \ldots, u_{k-1} \rangle$.

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": $|\,\mathrm{Vol}\,|\mathcal{P}_1(A; u_1) = \|u_1\|$, $|\,\mathrm{Vol}\,|\mathcal{P}_k(A; u_1, \ldots, u_k) = \|e_k\|\,|\,\mathrm{Vol}\,|\mathcal{P}_{k-1}(A; u_1, \ldots, u_{k-1})$. If $u_1, \ldots, u_n$ is a basis compatible with the orientation of V, the (oriented) volume of the parallelepiped is defined by

$$\mathrm{Vol}\,\mathcal{P}_n(A; u_1, \ldots, u_n) = |\,\mathrm{Vol}\,|\mathcal{P}_n(A; u_1, \ldots, u_n).$$

In the case of a non compatible basis we set

$$\mathrm{Vol}\,\mathcal{P}_n(A; u_1, \ldots, u_n) = -|\,\mathrm{Vol}\,|\mathcal{P}_n(A; u_1, \ldots, u_n).$$

Theorem. *Let* $\mathcal{Q} \subset \mathcal{E}_n$ *be a Euclidean subspace, and let* $(e_1, \ldots, e_k)$ *be its orthonormal basis. For arbitrary vectors* $u_1, \ldots, u_k \in Z(\mathcal{Q})$ *and* $A \in \mathcal{Q}$ *the following holds*

(1) $\mathrm{Vol}\,\mathcal{P}_k(A; u_1, \ldots, u_k) = \begin{vmatrix} u_1 \cdot e_1 & \ldots & u_k \cdot e_1 \\ \vdots & & \vdots \\ u_1 \cdot e_k & \ldots & u_k \cdot e_k \end{vmatrix}$

(2) $(\mathrm{Vol}\,\mathcal{P}_k(A; u_1, \ldots, u_k))^2 = \begin{vmatrix} u_1 \cdot u_1 & \ldots & u_k \cdot u_1 \\ \vdots & & \vdots \\ u_1 \cdot u_k & \ldots & u_k \cdot u_k \end{vmatrix}$

PROOF. The matrix

$$A = \begin{pmatrix} u_1 \cdot e_1 & \ldots & u_k \cdot e_1 \\ \vdots & & \vdots \\ u_1 \cdot e_k & \ldots & u_k \cdot e_k \end{pmatrix}$$

has the coordinates of the vectors $u_1, \ldots, u_k$ in the chosen basis in columns, and

$$|A|^2 = |A||A| = |A^T||A| = |A^T A|$$
$$= \begin{vmatrix} u_1 \cdot u_1 & \ldots & u_k \cdot u_1 \\ \vdots & & \vdots \\ u_1 \cdot u_k & \ldots & u_k \cdot u_k \end{vmatrix}.$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 0 & | & 0 \\ 0 & 1 & 0 & 0 & | & -1 \\ 0 & 0 & 1 & 0 & | & -1 \\ 0 & 0 & 0 & 1 & | & 2 \end{pmatrix}.$$

The solution is $(p_1, q_1, p_2, q_2) = (0, -1, -1, 2)$. Consequently

$$X_1 - X_2 = (-3, 2, -5, -7, -3) - u_2 + v_1 - 2v_2$$
$$= (-3, 4, -2, -4, 2).$$

The length of the vector $(-3, 4, -2, -4, 2)$ equals the distance between the planes σ_1, σ_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 σ_1, σ_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

$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 \\ -2 & 0 & 1 & 1 & 0 \\ 2 & 2 & 4 & 0 & 3 \\ 2 & 0 & 0 & -2 & -1 \end{pmatrix} \sim \cdots$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 0 & 3/2 \\ 0 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix},$$

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 σ_1 and σ_2 equals the size of the orthogonal projection of the vector (the difference of an arbitrary point in σ_1 and an arbitrary point in σ_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 σ_1 and σ_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 ϱ_1 and ϱ_2. The second method however reveals points where the distance can be measured (a pair of points in which the planes are the

Hence if (1) holds, then also (2) holds.

Directly from the definition, the unoriented volume equals the product

$$|\operatorname{Vol}|\mathcal{P}_k(A; u_1, \ldots, u_k) = \|v_1\| \|v_2\| \ldots \|v_k\|,$$

where $v_1 = u_1$, $v_2 = u_2 + a_1^2 v_1, \ldots, v_k = u_k + a_1^k v_1 + \cdots + a_{k-1}^k v_{k-1}$ is the result of the Gramm-Schmidt orthogonalization. Thus

$$(\operatorname{Vol} \mathcal{P}_k(A; u_1, \ldots, u_k))^2 = \begin{vmatrix} v_1 \cdot v_1 & 0 & \ldots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \ldots & v_k \cdot v_k \end{vmatrix}$$
$$= \begin{vmatrix} v_1 \cdot v_1 & \ldots & v_k \cdot v_1 \\ \vdots & & \vdots \\ v_1 \cdot v_k & \ldots & v_k \cdot v_k \end{vmatrix}.$$

Denote by B the matrix whose columns are formed by the coordinates of vectors $v_1, \ldots, v_k$ in the orthonormal basis $\underline{e}$. Since $v_1, \ldots, v_k$ have arisen from $u_1, \ldots, u_k$ 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 $\operatorname{Vol} \mathcal{P}_k(A; u_1, \ldots, u_k) = \pm |A|$. The resulting volume is zero if the vectors $u_1, \ldots, u_k$ are linearly dependent. Provided they are independent, the sign of the determinant is positive if and only if the basis $u_1, \ldots, u_k$ defines the same orientation as the basis $\underline{e}$. $\quad \square$

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 then the dimension of the whole space.

We formulate the following important geometric consequence:

4.1.23. Corollary. *For each linear map $\varphi : V \to V$ on a Euclidean 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.*

4.1.24. Outer product and cross product of vectors. The previous considerations are closely related to the tensor product of vectors. We do not go further in this technically more complicated topic. But we do mention the outer product $n = \dim V$ of vectors $u_1, \ldots, u_n \in V$.

closest). Find such points in the case of planes ϱ_1, ϱ_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

$$\begin{array}{rcrcrcrcr} 2t_1 & & & & & = & -5, \\ & 2s_1 & & + & 5s_2 & = & 1, \\ & & -4t_2 & - & s_2 & = & -2, \\ & -5s_1 & - & t_2 & - & 13s_2 & = & -1 \end{array}$$

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 ϱ_1, ϱ_2) both of which are given by $\| X_1 - X_2 \| = \| (0,0,0,3,0) \| = 3$. □

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 ϱ. Its direction is orthogonal to the direction of ϱ, for vectors (a,b,c,d) within its direction we get linear equation system

$$(a,b,c,d) \cdot (1,2,-1,-2) = 0 \quad \equiv \quad a + 2b - c - 2d = 0$$
$$(a,b,c,d) \cdot (1,0,0,1) = 0 \quad \equiv \quad a + d = 0.$$

Let $(u_{1j}, \ldots, u_{nj})^T$ 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. Its value is called the *outer product* of the vectors $u_1, \ldots, u_n$, and is denoted by $[u_1, \ldots, u_n]$. Hence the outer product is the oriented product of the corresponding parallelepiped, see 4.1.22.

Several useful properties of the outer product follow directly from the definition

(1) The map $(u_1, \ldots, u_n) \mapsto [u_1, \ldots, u_n]$ 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 $u_1, \ldots, u_n$ are linearly dependent.
(3) The vectors $u_1, \ldots, u_n$ 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 $u_1, \ldots, u_{n-1} \in V$. 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 $u_1, \ldots, u_{n-1}$. For each vector $w \in V$

$$\langle v, w \rangle = [u_1, \ldots, u_{n-1}, w].$$

We denote the cross product by $v = u_1 \times \ldots \times u_{n-1}$.

If the coordinates of the vectors in an orthonormal basis are $v = (y_1, \ldots, y_n)^T$, $w = (x_1, \ldots, x_n)^T$ and $u_j = (u_{1j}, \ldots u_{nj})^T$, then the definition can be expressed as

$$y_1 x_1 + \cdots + y_n x_n = \begin{vmatrix} u_{11} & \cdots & u_{1(n-1)} & x_1 \\ \vdots & & \vdots & \vdots \\ u_{n1} & \cdots & u_{n(n-1)} & x_n. \end{vmatrix}$$

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 $v = u_1 \times \ldots \times u_{n-1}$*

(1) $v \in \langle u_1, \ldots, u_{n-1} \rangle^\perp$
(2) v is nonzero if and only if the vectors $u_1, \ldots, u_{n-1}$ are linearly independent,
(3) the length $\|v\|$ of the cross product equals the absolute value of the volume of parallelepiped $\mathcal{P}(0; u_1, \ldots, u_{n-1})$,
(4) $(u_1, \ldots, u_{n-1}, v)$ 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 solution is the two-dimensional vector space $\langle(0,1,2,0),(-1,0,-3,1)\rangle$. The plane τ orthogonal to ϱ 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 ϱ, 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 τ). $\qquad\square$

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 $30°$.

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 $30°$. The rotation matrix for the angle $30°$ is

$$\begin{pmatrix} \cos 30° & -\sin 30° \\ \sin 30° & \cos 30° \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}.$$

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

$$[1,2] + \left(\frac{\sqrt{3}}{2} - \frac{1}{2}, \frac{\sqrt{3}}{2} + \frac{1}{2} \right) t.$$

$\qquad\square$

4.B.22. An octahedron has eight faces consisting of equilateral triangles. Determine $\cos\alpha$, where α 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

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 $u_1, \ldots, u_{n-1}$ are linearly dependent, then (3) also holds. Suppose the vectors are linearly independent. Let v be their cross product, and choose an orthonormal basis $(e_1, \ldots, e_{n-1})$ of the space $\langle u_1, \ldots, u_{n-1}\rangle$. It follows from what is proved that there exists a multiple $(1/\alpha)v, 0 \neq \alpha \in \mathbb{R}$, such that $(e_1, \ldots, e_k, (1/\alpha)v)$ is an orthonormal basis of V. The coordinates of the vectors in this basis are

$$u_j = (u_{1j}, \ldots, u_{(n-1)j}, 0)^T, \quad v = (0, \ldots, 0, \alpha)^T.$$

So the outer product $[u_1, \ldots, u_{n-1}, v]$ equals (see the definition of cross product)

$$[u_1, \ldots, u_{n-1}, v] = \begin{vmatrix} u_{11} & \cdots & u_{1(n-1)} & 0 \\ \vdots & & \vdots & \vdots \\ u_{(n-1)1} & \cdots & u_{(n-1)(n-1)} & 0 \\ 0 & \cdots & 0 & \alpha \end{vmatrix}$$
$$= \langle v, v \rangle = \alpha^2.$$

By expanding the determinant along the last column,

$$\alpha^2 = \alpha \operatorname{Vol} \mathcal{P}(0; u_1, \ldots, u_{n-1}).$$

Both the remaining two claims follow from the proposition below. $\qquad\square$

In technical applications in $\mathbb{R}^3$, the cross product is often used. It assigns a vector to any pair of vectors.

4.1.25. Affine and Euclidean properties. Now we can consider which properties are related to the affine structure of the space and which properties we really need in the difference space.

All Euclidean transformations, (bijective affine maps) which preserve the distance between points, preserve also all objects we have studied. Moreover they preserve unoriented angles, unoriented volumes, angle between subspaces etc. If we want them to preserve also oriented angles, cross products, volumes, then we must also assume that the transformations preserve the orientation.

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

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 ythe 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 $\alpha \doteq 132°$. $\qquad\square$

4.B.23. In Euclidean space $\mathbb{R}^5$ determine the angle φ 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) $U : [0,1,1,0,0] + t\,(0,0,0,1,-1),\ t \in \mathbb{R}$,
   $V : [1,0,1,1,1] + r\,(1,-1,2,1,0) + s\,(0,1,3,2,0)$
   $\qquad + p\,(1,0,0,1,0) + q\,(1,3,1,0,0),$
   $r,s,p,q \in \mathbb{R}$;

(e) $U : [0,2,5,0,0] + t\,(2,1,3,5,3) + s\,(0,3,1,4,-2)$
   $\qquad + r\,(1,2,4,0,3),\ t,s,r \in \mathbb{R}$,
   $V : [0,0,0,0,0] + p\,(-1,1,1,-5,0)$
   $\qquad + q\,(1,5,1,13,-4),\ p,q \in \mathbb{R}$;

(f) $U : [1,1,1,1,1] + t\,(1,0,1,1,1)$
   $\qquad + s\,(1,0,0,1,1),\ t,s \in \mathbb{R}$,
   $V : [1,1,1,1,1] + p\,(1,1,1,1,1) + q\,(1,1,0,1,1)$
   $\qquad + r\,(1,1,0,1,0), p,q,r \in \mathbb{R}$.

Solution. Recall that the angle between affine subspaces is the same as the angle between vector spaces associated to

transformations, since the determinant of the corresponding matrix changes by the same multiple.

These affine properties can be used in the plane to prove geometric statements. For instance, 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 case of an equilateral triangle. Then this property holds for all triangles. Think about this argument!

2. Geometry of quadratic forms

After straight lines, 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, therefore we begin with Euclidean geometry.

4.2.1. Quadrics in $\mathcal{E}_n$. In analogy with the equations of conic sections in plane, we start with objects in Euclidean point spaces. These are defined in a given orthonormal basis by quadratic equations, and are known as *quadrics*.

Choose a fixed Cartesian coordinate system in $\mathcal{E}_n$. This is a point and an orthonormal basis of the difference space. Consider a general quadratic equation for the coordinates $(x_1,\ldots,x_n)^T$ of a point $A \in \mathcal{E}_n$

$$(1) \qquad \sum_{i,j=1}^{n} a_{ij}x_ix_j + \sum_{i=1}^{n} 2a_ix_i + a = 0,$$

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 Euclidean subspace.

Notice that the equation (1) keeps the same shape under every Euclidean (or affine) coordinate transformation, i.e., it splits again into quadratic, linear and constant parts.

4.2.2. Quadratic forms. Begin the discussion of equation (1) with its quadratic part, i.e. bilinear symmetric form $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Similarly, think of a general symmetric bilinear form on an arbitrary vector space.

For an arbitrary basis on this vector space, the value $f(x)$ on vector $x = x_1e_1 + \cdots + x_ne_n$ is given by the equation

$$(1) \qquad f(x) = F(x,x) = \sum_{i,j} x_ix_j F(e_i,e_j) = x^T \cdot A \cdot x$$

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

$$\cos\varphi = \frac{|\,(1,0,2,-2,1)\cdot(2,0,-2,1,-1)\,|}{||\,(1,0,2,-2,1)\,||\cdot||\,(2,0,-2,1,-1)\,||} = \frac{5}{\sqrt{10}\cdot\sqrt{10}}.$$

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

$$\cos\psi = \frac{(2,0,0,2,1)\cdot(1,1,1,0,1)}{||\,(2,0,0,2,1)\,||\cdot||\,(1,1,1,0,1)\,||} = \frac{3}{3\cdot2}.$$

Notice that $\varphi = \pi/2 - \psi = \pi/6$, because φ is complement to ψ.

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 φ 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

$$u = (0,0,0,1,-1), \quad v_1 = (1,-1,2,1,0),$$

$$v_2 = (0,1,3,2,0),\ v_3 = (1,0,0,1,0),\ v_4 = (1,3,1,0,0)$$

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

$$\langle\, p_u - u, v_1 \,\rangle = 0, \quad \langle\, p_u - u, v_2 \,\rangle = 0,$$

$$\langle\, p_u - u, v_3 \,\rangle = 0, \quad \langle\, p_u - u, v_4 \,\rangle = 0.$$

Substituting for p_u gives the linear equation system

$$\begin{array}{rcrcrcrcl}
7a &+& 7b &+& 2c & & &=& 1, \\
7a &+& 14b &+& 2c &+& 6d &=& 2, \\
2a &+& 2b &+& 2c &+& d &=& 1, \\
 & & 6b &+& c &+& 11d &=& 0.
\end{array}$$

The solution is $(a,b,c,d) = (-8/19, 7/19, 13/19, -5/19)$.

$$(1) \qquad \cos\varphi = \frac{||\,p_u\,||}{||\,u\,||}$$

and so

$$p_u = -\frac{8}{19}v_1 + \frac{7}{19}v_2 + \frac{13}{19}v_3 - \frac{5}{19}v_4 = (0,0,0,1,0),$$

$$\cos\varphi = \frac{||\,p_u\,||}{||\,u\,||} = \frac{||\,(0,0,0,1,0)\,||}{||\,(0,0,0,1,-1)\,||} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}.$$

Hence $\varphi = \pi/4$.

Case (e). Determine the intersection of the vector subspaces associated with the given affine subspaces. The

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 from above for the value of the form in terms of the chosen coordinates is called the *analytic formula* for the form.

In general, by a quadratic form is meant the restriction $f(x)$ of a symmetric bilinear form $F(x,y)$ to arguments of the type (x,x). Evidently, the whole bilinear form F can be reconstructed from the values $f(x)$ since

$$f(x+y) = F(x+y, x+y) = f(x) + f(y) + 2F(x,y).$$

If we change the basis e_i 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

$$f(x) = (S\cdot x')^T \cdot A \cdot (S\cdot x') = (x')^T \cdot (S^T \cdot A \cdot S) \cdot x'.$$

Assume now that the vector space is equipped with a scalar product. Then the previous computation can be formulated as follows. 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 $F(\ ,u)$. 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 (1) displayed above, φ is the linear map with symmetric matrix A. Hence it is selfadjoint.

On the other hand, each symmetric map φ 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. $\qquad\square$

It is immediate that for each quadratic form f there exists an orthonormal basis of the difference space in which f has a diagonal matrix. The values on the diagonal are determined uniquely up to their order.

Due to the identification of quadratic forms with linear maps, the *rank of the quadratic form* can be defined as the rank of its matrix in any basis. The rank equals the dimension of the image of the corresponding map φ).

4.2.3. Classification of quadrics. We return to the equation (1). The above results enable us to rewrite this equation 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.

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

$$t\,(2, 1, 3, 5, 3) + s\ (0, 3, 1, 4, -2) + r\,(1, 2, 4, 0, 3)$$
$$= p\,(-1, 1, 1, -5, 0) + q\,(1, 5, 1, 13, -4).$$

It is a homogeneous linear equation system. It is solved in matrix form (order of variables is t, s, r, p, q)

$$\begin{pmatrix} 2 & 0 & 1 & 1 & -1 \\ 1 & 3 & 2 & -1 & -5 \\ 3 & 1 & 4 & -1 & -1 \\ 5 & 4 & 0 & 5 & -13 \\ 3 & -2 & 3 & 0 & 4 \end{pmatrix} \sim \cdots$$

$$\sim \begin{pmatrix} 1 & 3 & 2 & -1 & -5 \\ 0 & 2 & 1 & -1 & -3 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

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 φ is defined as the angle between the subspaces

$$Z(U) \cap (Z(U) \cap Z(V))^\perp \quad \text{a} \quad Z(V) \cap (Z(U) \cap Z(V))^\perp.$$

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

$$\sum_{i=1}^{n} \lambda_i (x_i - p_i)^2\ \ + \sum_{j=1}^{n} b_j x_j + c = 0.$$

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 φ 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 φ so that the corresponding linear form is a multiple of the first term of the dual basis. Hence the final formula reads

$$\sum_{i=1}^{k} \lambda_i y_i^2 + b y_{k+1} + c = 0,$$

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. $c \in \mathbb{R}$ may be nonzero only if $b = 0$. The resulting equations are called the *canonical analytic formulas* for quadrics.

4.2.4. The case of $\mathcal{E}_2$. As an example of the previous procedure, we discuss the simplest case of a nontrivial dimension, namely dimension two. The original equation has the form

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_1 x + a_2 y + a = 0.$$

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):

$$a_{11}x^2 + a_{22}y^2 + a_1 x + a_2 y + a = 0$$

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:

$0 = x^2/a^2 + y^2/b^2 + 1$	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

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

$$\langle\, p_u - u, (1,1,0,1,1) \rangle = 0, \ \langle\, p_u - u, (1,1,0,1,0) \rangle = 0,$$

which leads to the linear equation system

$$\begin{array}{rcrcl} 4a &+& 3b &=& 3, \\ 3a &+& 3b &=& 2 \end{array}$$

with the unique solution $a = 1, b = -1/3$. Thus

$$p_u = \left(\tfrac{2}{3}, \tfrac{2}{3}, 0, \tfrac{2}{3}, 1 \right).$$

From (1) it follows that

$$\cos \varphi = \frac{|| (2/3, 2/3, 0, 2/3, 1) ||}{|| (1,0,0,1,1) ||} = \frac{\sqrt{7}}{3}. \quad \varphi \doteq 0.49 \ (\approx 28\,^{\circ}).$$

$\square$

C. Geometry of quadratic forms

4.C.1. Determine the polar basis of the form $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = 3x_1^2 + 2x_1 x_2 + x_2^2 + 4x_2 x_3 + 6x_3^2$.

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 4.2.5), perform the following operations

$$\begin{aligned} f(x_1, x_2, x_3) &= \frac{1}{3}(3x_1 + x_2)^2 + \frac{2}{3}x_2^2 + 4x_2 x_3 + 6x_3^2 \\ &= \frac{1}{3}y_1^2 + \frac{3}{2}\left(\frac{2}{3}y_2 + 2y_3\right)^2 \\ &= \frac{1}{3}z_1^2 + \frac{3}{2}z_2^2. \end{aligned}$$

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

$$T = \begin{pmatrix} 3 & 1 & 0 \\ 0 & \frac{2}{3} & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

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 nondegenerate directions.

4.2.5. Affine point of view. In the previous two paragraphs, we searched for essential properties and standardized analytical descriptions of objects defined in Euclidean spaces by quadratic equations. We sought the simplest equations which can be obtained by a suitable choice of coordinates. A geometric formulation of the result is that for two different quadrics, given in different Cartesian coordinates, there exists a *Euclidean transformation* on $\mathcal{E}_n$ (that is, an affine bijective map preserving lengths) if and only if the above algorithm leads to the same analytic formulas, up to the order of coordinates. Moreover, the Cartesian coordinates in which the objects are given by the resulting canonical formulas, can be obtained directly. Hence the explicit expression of the corresponding coordinate transformation is also obtained. It is always a composition of a translation, rotation and reflection with respect to a hyperplane.

Of course, we may ask to what extent we can do the same in affine spaces, where we can choose any coordinate system. For example, in the plane we cannot distinguish the circle from the ellipse. On the other hand, we can distinguish from the hyperbola and between all other types of conics. In particular, all hyperbolas merge into one etc. We postpone discussion of this issue to the third part of this chapter, except for the case of quadratic forms.

Consider a quadratic form f on a vector space V and its analytic formula $f(u) = x^T A x$ with respect to a chosen basis on V. Then for the vector $u = x_1 u_1 + \cdots + x_n u_n$, the form f can be written as

$$f(x_1, \ldots, x_n) = \sum_{ij} a_{ij} x_i x_j,$$

It is already shown that A is diagonal for a suitable choice of basis. In other words that $F(u_i, u_j) = 0$ for $i \neq j$ for a suitable symmetric form F. Each such basis is called the *polar basis* of the quadratic form f. A scalar product can always be chosen for such a purpose. 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, there is relevant information to be found 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 V be a real vector space of dimension n, $f : V \to \mathbb{R}$ a quadratic form. Then there exist a polar basis for f on V.*

255

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 $((\frac{1}{3}, 0, 0), (-\frac{1}{2}, \frac{3}{2}, 0), (1, -3, 1))$. $\square$

4.C.2. Determine the polar basis of the form $f : \mathbb{R}^3 \to \mathbb{R}^3$. $f(x_1, x_2, x_3) = 2x_1 x_3 + x_2^2$.

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 $z_1 = y_1$, $z_2 = y_2$, $z_3 = y_3 - y_2$.

$$f(x_1, x_2, x_3) = z_1^2 + 2z_2(z_3 + z_2) = z_1^2 + \frac{1}{2}(2z_2 + z_3)^2 - \frac{1}{2}z_3^2.$$

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 $f : \mathbb{R}^3 \to \mathbb{R}$, which in the standard basis is defined as

$$f(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3.$$

PROOF. (1) Let A be the matrix of f in basis $\underline{u} = (u_1, \ldots, u_n)$ on V, and assume $a_{11} \neq 0$. Then we may write

$$f(x_1, \ldots, x_n) = a_{11}x_1^2 + 2a_{12}x_1 x_2 + \cdots + a_{22}x_2^2 + \ldots$$
$$= a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)^2$$
$$+ \text{ terms not containing } x_1.$$

Hence we can transform the coordinates (i.e. change the basis) such that in the new coordinates

$$x_1' = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n, \; x_2' = x_2, \ldots, x_n' = x_n.$$

This corresponds to the new basis

$$v_1 = a_{11}^{-1}u_1, \; v_2 = u_2 - a_{11}^{-1}a_{12}u_1, \ldots, v_n = u_n - a_{11}^{-1}a_{1n}u_1.$$

(As an exercise, compute the transformation matrix). In the new basis the corresponding symmetric bilinear form satisfies $g(v_1, v_i) = 0$ for all $i > 1$ (compute it!). Thus f has the 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 $g(v_1, v_1) = a_{11}$.

(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 $v_j = u_j + u_k +$ we keep the other vector of basis constant (i.e. $x_k' = x_k - x_j$, the other remain constant). 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). $\square$

4.2.6. 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 an affine 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*

Solution. By an application of the Lagrange algorithm:

$$f(x_1, x_2, x_3) = 2x_1x_2 + x_2x_3$$

substitution $y_2 = x_2 - x_1$, $y_1 = x_1$, $y_3 = x_3$

$$= 2x_1(x_1 + y_2) + (x_1 + y_2)x_3$$
$$= 2x_1^2 + 2x_1y_2 + x_1x_3 + y_2x_3$$
$$= \frac{1}{2}(2x_1 + y_2 + \frac{1}{2}x_3)^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3$$

substitution $y_1 = 2x_1 + y_2 + \frac{1}{2}x_3$

$$= \frac{1}{2}y_1^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3$$
$$= \frac{1}{2}y_1^2 - 2(\frac{1}{2}y_2 - \frac{1}{2}x_3)^2 + \frac{3}{8}x_3^2$$

substitution $y_3 = \frac{1}{2}y_2 - \frac{1}{2}x_3$

$$= \frac{1}{2}y_1^2 - 2y_3^2 + \frac{3}{8}x_3^2.$$

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), $\{(1/3, 1/3, 0), (-2/3, 4/3, 0), (-1/2, 1/2, 1)\}$. □

4.C.4. Determine the type of conic section defined by

$$3x_1^2 - 3x_1x_2 + x_2 - 1 = 0.$$

independent linear forms $\varphi_1, \ldots, \varphi_r \in V^*$ such that $0 \le p \le r$ and

$$f(u) = (\varphi_1(u))^2 + \cdots + (\varphi_p(u))^2 - (\varphi_{p+1}(u))^2 - \cdots - (\varphi_r(u))^2.$$

Otherwise put, there exists a polar basis, in which f has an analytic formula

$$f(x_1, \ldots, x_n) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_r^2.$$

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 square, $f(x_1, \ldots, x_n) = \lambda_1 x_1^2 + \cdots + \lambda_r x_r^2$, $\lambda_i \ne 0$, in a basis on V. Assume moreover that the first p coefficients λ_i are positive. Then the transformation $y_1 = \sqrt{\lambda_1}x_1, \ldots, y_p = \sqrt{\lambda_p}x_p, y_{p+1} = \sqrt{-\lambda_{p+1}}x_{p+1}, \ldots, y_r = \sqrt{-\lambda_r}x_r, y_{r+1} = x_{r+1}, \ldots, y_n = x_n$ yields the desired formula. The forms φ_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.

$$f(x_1, \ldots, x_n) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_r^2$$
$$f(y_1, \ldots, y_n) = y_1^2 + \cdots + y_q^2 - y_{q+1}^2 - \cdots - y_r^2.$$

Denote the subspace generated by the first p vectors of the first basis by $P = \langle u_1, \ldots, u_p \rangle$, and similarly $Q = \langle v_{q+1}, \ldots, v_n \rangle$. Then for each $u \in P$, $f(u) > 0$ while for $v \in Q$ $f(v) \le 0$. Hence necessarily $P \cap Q = \{0\}$, and therefore $\dim P + \dim Q \le n$. Hence $p + (n - q) \le n$, so that $p \le q$. By interchanging the subspaces, $q \le p$, 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. . □

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. A quadratic form f on a real vector space V is called

(1) *positive definite* if $f(u) > 0$ for all vectors $u \ne 0$,
(2) *positive semidefinite* if $f(u) \ge 0$ for all vectors $u \in V$,
(3) *negative definite* if $f(u) < 0$ for all vectors $u \ne 0$,
(4) *negative semidefinite* if $f(u) \le 0$ for all vectors $u \in V$,
(5) *indefinite* if $f(u) > 0$ and $f(v) < 0$ for two vectors $u, v \in V$.

Solution. Complete the squares:

$$3x_1^2 - 3x_1x_2 + x_2 - 1 = \frac{1}{3}(3x_1 - \frac{3}{2}x_2)^2 - \frac{3}{4}x_2^2 + x_2 - 1$$
$$= \frac{1}{3}y_1^2 - \frac{4}{3}(\frac{3}{4}x_2 - \frac{1}{2})^2 + \frac{1}{3} - 1$$
$$= \frac{1}{2}y_1^2 - \frac{4}{3}y_2^2 - \frac{2}{3}.$$

According to the list 4.2.4, the given conic section is a hyperbola. □

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)

$$-\bar{x}^2 + \bar{y}^2 + \bar{z}^2 - 1 = 0.$$

□

We can tell the type of the conic section without transforming its equation to the form listed in 4.2.4. 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 $\quad \Delta = \det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{32} & a_{33} \end{vmatrix}$ 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 invariant to the coordinate transformation. Denote $X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$

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.

4.2.7. Theorem (Sylvester criterion). *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 $V_k = \langle u_1, \ldots, u_k \rangle$ are positive definite forms f_k again, and the corresponding matrices in the bases $u_1, \ldots, u_k$ 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 proposition 4.2.5, the matrix has one's on the diagonal:

$$T = \begin{pmatrix} 1 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{n2}}{a_{11}} \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \end{pmatrix}.$$

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 $P_k = \langle u_1, \ldots, u_k \rangle$ from basis $(u_1, \ldots, u_k)$ to basis $(v_1 \ldots, v_k)$. 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 $A_k = T_k^T A'_k (T_k)^{-1}$, 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. Thus, we have proved:

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*

$$f(x_1, \ldots, x_n) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_r x_r^2$$

is obtained where r is the rank of the form f, $\lambda_1, \ldots, \lambda_r \neq 0$ and for the leading principal submatrices of the (former) matrix A of quadratic form f, $|A_k| = \lambda_1 \lambda_2 \ldots \lambda_k$, $k \leq r$.

and denote A as the matrix of the quadratic form. Then the corresponding conic section has equation $X^T A X = 0$. The 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)
$$
X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha & c_1 \\ \sin\alpha & \cos\alpha & c_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = M X'.
$$

Put $X = M X'$ into the conic section equation to obtain the equation in new coordinates

$$
\begin{aligned}
X^T A X &= 0 \\
(M X')^T A (M X') &= 0 \\
X'^T M^T A M X' &= 0.
\end{aligned}
$$

Denote by A' the matrix of the quadratic form in new coordinates. Then $A' = M^T A M$, where matrix $M = \begin{pmatrix} \cos\alpha & -\sin\alpha & c_1 \\ \sin\alpha & \cos\alpha & c_2 \\ 0 & 0 & 1 \end{pmatrix}$ 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. $\square$

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 $|A_1| \neq 0, \ldots, |A_r| \neq 0$. Then there exists a polar basis in which f has the analytic formula*

$$
f(x_1, \ldots, x_n) = |A_1| x_1^2 + \frac{|A_2|}{|A_1|} x_2^2 + \cdots + \frac{|A_r|}{|A_{r-1}|} x_r^2.
$$

Hence if all leading principal minors are positive, then f is positive definite by the Jacobi theorem and the Sylvester criterion is proved. $\square$

3. 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.

4.3.1. 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 $\mathcal{R}^3$, 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.

4.C.7. Determine the type of conic section $x^2 - 4xy - 5y^2 + 2x + 4y + 3 = 0$.

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. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

4.C.8. Determine the equation and type of conic section passing through the points

$$[-2, -4], \quad [8, -4], \quad [0, -2], \quad [0, -6], \quad [6, -2].$$

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

$$
\begin{pmatrix}
4 & 16 & 16 & -2 & -4 & 1 \\
64 & 16 & -64 & 8 & -4 & 1 \\
0 & 4 & 0 & 0 & -2 & 1 \\
0 & 36 & 0 & 0 & -6 & 1 \\
36 & 4 & -24 & 6 & -2 & 1
\end{pmatrix} \sim \cdots
$$

$$
\sim
\begin{pmatrix}
4 & 16 & 16 & -2 & -4 & 1 \\
0 & 4 & 0 & 0 & -2 & 1 \\
0 & 0 & 64 & -8 & 12 & -9 \\
0 & 0 & 0 & 24 & -36 & 27 \\
0 & 0 & 0 & 0 & 3 & -2
\end{pmatrix} \sim \cdots
$$

$$
\sim
\begin{pmatrix}
48 & 0 & 0 & 0 & 0 & -1 \\
0 & 12 & 0 & 0 & 0 & -1 \\
0 & 0 & 64 & 0 & 0 & 0 \\
0 & 0 & 0 & 24 & 0 & 3 \\
0 & 0 & 0 & 0 & 3 & -2
\end{pmatrix}.
$$

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

$$(x-3)^2 + 4(y+4)^2 - 25 = 0,$$

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* $\mathbb{R}^3$.

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.

4.3.2. 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 σ 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 σ.

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 affine coordinates.

4.3.3. Projective spaces and transformations. In a natural way one can generalize the procedure in the affine plane for each finite dimension.

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

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 $[3, -4]$. □

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 (4.2.4) 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.

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

$$\mathcal{P}(V) = \{P \subset V; \ P \subset V, \dim V = 1\}.$$

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 projective coordinates, we fix one of them to be one. Thus we exclude all points of the projective extension 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.

4.3.4. Perspective projection. The advantages of projective geometry show up well in the case of perspective projection $\mathbb{R}^3 \to \mathbb{R}^2$. 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 $Z = f > 0$.

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 $\mathbb{R}^3 \subset \mathcal{P}_3$ which we 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.

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.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

$$|PF_1|^2 = (a\cos\theta - ae)^2 + b^2\sin^2\theta$$
$$= a^2\cos^2\theta - 2a^2 e\cos\theta + a^2 e^2 + a^2(1 - e^2)\sin^2\theta$$
$$= a^2[-2e\cos\theta + e^2 - e^2(1 - \cos^2\theta)]$$
$$= a^2(1 - e\cos\theta)^2.$$

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 $|F_1 X| + |F_2 X| = 2a$. 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.

4.3.5. Affine and projective transformations. Each injective 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 $T : \mathcal{P}(V_1) \rightarrow \mathcal{P}(V_2)$. Such maps are called *projective maps.* In the literature, the notion *collineation* is used if this map is invertible.

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.

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.

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:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ b_1 & a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots & \\ b_n & a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

where $b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$ and $A = (a_{ij})$ is an invertible matrix of dimension n. The action of such a matrix on the vector $(1, x_1, \ldots, x_n)$ 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.

4.3.6. 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:
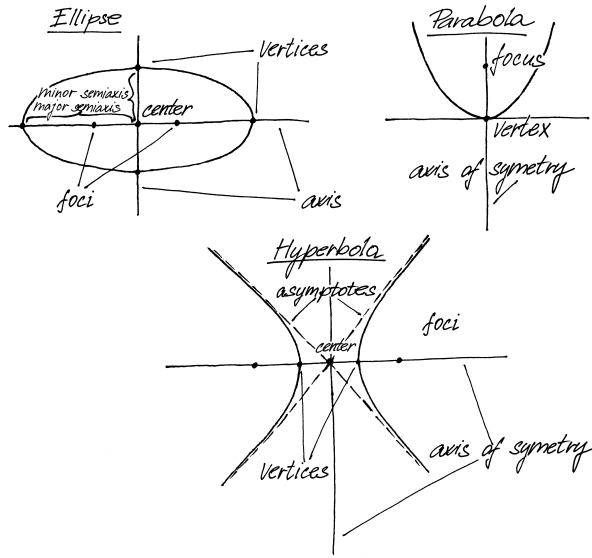
Similarly, the *hyperbola foci* are the points F_1, F_2, which satisfy $||F_2X| - |F_1X|| = 2a$ for an arbitrary X on the hyperbola. You can check this in same the way as above for the ellipse, with $F_1 = [ae, 0]$, $F_2 = [-ae, 0]$, $ae = \sqrt{a^2 + b^2}$.

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 $[1, 0]$. □

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 θ. The tangent line to the ellipse at T has equation

$$y - b\sin\theta = -\frac{b\cos\theta}{a\sin\theta}(x - a\cos\theta).$$

$$a(\sin\theta)(y - b\sin\theta) = -b(\cos\theta)(x - a\cos\theta).$$

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

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

$$z = c_1 u + c_2 v + c_3 w, \quad z' = c_1' u' + c_2' v' + c_3' w',$$

where all six coefficients must be nonzero, otherwise there exist three points not in general position.

Now choose new arithmetic representatives $\tilde{u} = c_1 u$, $\tilde{v} = c_2 v$ and $\tilde{w} = c_3 w$ of points A, B and C respectively. Similarly $\tilde{u}' = c_1' u'$, $\tilde{v}' = c_2' v'$ and $\tilde{w}' = c_3' w'$ for points A', B' and C'. This choice defines an unique linear map φ 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 φ 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. □

4.3.7. 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. 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

$$w = t_1 x + s_1 y, \quad z = t_2 x + s_2 y.$$

φ to get

$$D_1^2 = a^2(1 - e\cos\theta)^2 \left[\frac{b^2}{a^2\sin^2\theta + b^2\cos^2\theta}\right]$$

$$= (1 - e\cos\theta)^2 \left[\frac{b^2}{\sin^2\theta + (1 - e^2)\cos^2\theta}\right]$$

$$= (1 - e\cos\theta)^2 \left[\frac{b^2}{1 - e^2\cos^2\theta}\right]$$

$$= b^2(1 - e\cos\theta) \left[\frac{1}{1 + e\cos\theta}\right]$$

Since D_2 is the same as D_1 with e replaced by $-e$, it follows that $D_1 D_2 = b^2$. $\qquad\square$

Solution. (Alternative). Consider the polar basis. The ellipse matrix has diagonal shape $\mathrm{diag}(\frac{1}{a^2}, \frac{1}{b^2}, -1)$ 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 $a^2 e^2 = a^2 - b^2$ 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. $\qquad\square$

4.C.13. Projective approach to conic section.

Projective space gives an ability to approach the conic section from a new perspective (compare with 4.3.11). We can understand conic sections in $\mathcal{E}_2$ defined by the quadratic form

$$f(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33}$$

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 $f(v) = v^T A v$, where v is a column vector with coordinates (x, y, z) and matrix A is symmetric matrix (a_{ij}). By theorem 4.2.6, 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

Define the *cross-ratio of four points* (A, B, C, D) as

$$\rho = \frac{s_1}{t_1}\frac{t_2}{s_2}.$$

The definition is valid, since although the vectors x and y are determined up to a scalar multiple, these multiples cancel out in the definition.

Similarly, each projective transformation preserves cross-ratios. Indeed, if the transformation is given in arithmetic coordinates by a matrix A, we have images $A \cdot w = t_1 A \cdot x + t_2 A \cdot y$, 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.

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 projective transformation of the projective extension $\mathcal{P}\mathbb{R}^{n+1}$ of the former affine space $\mathbb{R}^n$. Thus these transformations also preserve cross-ratios.

4.3.8. Duality.
The projective hyperplanes in 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 α 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.

Of forms, the linear map defining a given collineation acts by the multiplication of row vectors from the right by the same matrix

$$\alpha = (\alpha_0, \dots, \alpha_n) \mapsto \alpha \cdot A.$$

The matrix of the dual map is 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^*_{alg}, up to the multiplication by the inverse of determinant, (see equation (1) on page 91,) we can work directly with the projective transformation of the space $\mathcal{P}(V^*)$ given by the matrix $(A^*_{\mathrm{alg}})^T$ (or without transposing if we multiply row vectors from the right).

Ellipse   Parabola   Hyperbola
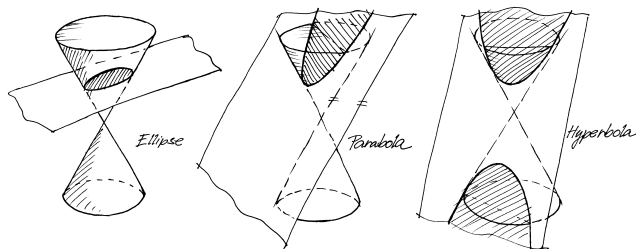
cone with the plane which has the equation $z = 1$ in the original basis. Immediately we obtain the conic section classification from 4.29., which corresponds to the intersecting cone in $\mathbb{R}^3$ with different planes. Non-degenerate sections are depicted. Degenerate sections are those which pass 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 $p^T A q = 0$.

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 $F(p, x) = p^T A x = 0$. 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 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 $\alpha, \beta \in \mathbb{R}$.

The projective point X belongs to the hyperplane α if the arithmetic coordinates satisfy $\alpha \cdot x = 0$. It still holds after acting with an arbitrary collineation, since

$$(\alpha \cdot A^{-1}) \cdot (A \cdot x) = \alpha \cdot x = 0.$$

4.3.9. 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 A which is mapped to itself. That is, $f(A) = A$. By the *fixed hyperplane* of collineation f is meant a hyperplane α 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* $A \in \mathcal{P}(V)$ is a set of all hyperplanes which contain the point A. For each point A 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 $f : \mathcal{P}(V) \to \mathcal{P}(V)$, 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 α 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 other than the identity has either exactly one center and exactly one axis, or it has neither a center nor an axis.*
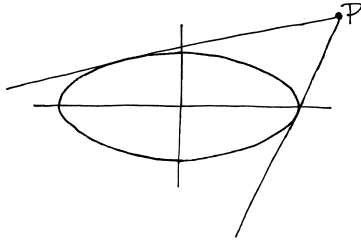
Because of the bilinearity and symmetry of F,

$$f(x) = F(x,x) = \alpha^2 F(p,p) + 2\alpha\beta F(p,q) + \beta^2 F(q,q) = 0.$$

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

$$\begin{pmatrix} U^T & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{A} & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} D & U^T a \\ a^T U & \alpha \end{pmatrix}$$

PROOF. Consider a collineation f on $\mathcal{P}\mathbb{R}^3$ and assume that it has two distinct centers A and B. Denote by ℓ the line given by these two centers, and choose a point X in the projective plane outside of ℓ. 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 ℓ are fixed. Hence each line different from ℓ has all points out of ℓ fixed and thus also its intersection with ℓ 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. □ check in detail!

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. picture missing!

4.3.10. 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*

$$Q = [AB'] \cap [BA'], \ R = [AC'] \cap [CA'], \ S = [BC'] \cap [CB'].$$

Then $\{Q, R, S\}$ are also collinear.

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 $\{T, A', B', C'\}$

$$b' = t + a', \quad c' = \lambda' t + a'$$

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_λ, v_μ, 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_λ and u_μ 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_λ a v_μ. □

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

$$(x_0, y_0, 1) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -p \\ 0 & -p & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = x_0 x - py - py_0 = 0$$

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

$$\frac{1}{\sqrt{p^2 + x_0^2}}(p, x_0).(0, 1) = \frac{x_0}{\sqrt{p^2 + x_0^2}}$$

Now we show that this is the same as the cosine of the angle between the tangent and line connecting the focus F$=[0, \frac{p}{2}]$,

for some real constants λ and λ'. 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,

$$s = q - r = t + a + a' - \lambda\lambda't - \lambda'a - \lambda a'$$
$$= (1 - \lambda')(t + a) + (1 - \lambda)(\lambda't + a')$$
$$= (1 - \lambda)(t + a') + (1 - \lambda')(\lambda t + a)$$

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. □

4.3.11. 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 (1), see page 252. 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 $(x_0, x_1, \ldots, x_n)$ 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 (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 4.2.6. 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 $x^2 + y^2 + z^2 = 0$

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,

$$\frac{1}{\sqrt{p^2 + x_0^2}} \frac{1}{\sqrt{x_0^2 + (y_0 - \frac{p}{2})^2}}(x_0 y_0 + \frac{px_0}{2})$$

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 $\frac{2at-0}{at-a} = \frac{2t}{(t^2-1)}$. If θ 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 θ.

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 D

- 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.

4.3.12. Affine classification of quadrics. For an affine classification we must restrict the projective transformations to those which preserve the line 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 α.

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 α we may for instance choose the tangent planes to unite the 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 α we obtain a more and more stretched ellipse until we get such a slope that α 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 infinite line 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. Find a parametric equation for the intersection of the following planes in $\mathbb{R}^3$:

$\sigma: 2x + 3y - z + 1 = 0$ a $\rho: x - 2y + 5 = 0.$

○

4.D.2. 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.3. Jarda is standing in $[-1, 1, 0]$ and has a stick of length 4. Can he simultaneously touch the lines p and q, where

$$p \ : \ [0, -1, 0] + t(1, 2, 1),$$
$$q \ : \ [3, 4, 8] + s(2, 1, 3)?$$

(The stick must pass through $[-1, 1, 0]$.)

○

4.D.4. 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.5. 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.6. 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.7. 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.8. 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. □

4.D.9. 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 θ.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

$\tan\varphi = \frac{b\cos\theta}{a\sin\theta}$ $3\cos\theta = a\sin\varphi$ where $a\sin\theta\tan\varphi = b\cos\theta$ $3\sin\theta = b\cos\varphi$
$9/a^2 \cos^2\theta + 9/b^2 \sin^2\theta = 1$. $9\cos^2\theta + 21\sin^2\theta = 10$. $12\sin^2\theta = 1$
$(y - \sqrt{35/2}) = -(x - \sqrt{55}/\sqrt{6})[\sqrt{33/7}]$ $\square$

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,

$$(3x_T)^2 + (7y_T)^2 = 100$$
$$3x_T^2 + 7y_T^2 = 30$$

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

$$\pm 3\sqrt{\frac{55}{6}}x \pm 7\sqrt{\frac{5}{14}}y - 30 = 0.$$ $\square$

4.D.10. 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 $a^2 e^2 = a^2 + b^2 = 4$, 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. $\square$

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 $a^2 e^2 = a^2 + b^2 = 4$, 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 $x^2 - \frac{y^2}{3} = 1$. $\square$

4.D.11. 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 $(36c)^2 - 4.32.(9c^2 + 1) = 0$, and $c = \pm \frac{2\sqrt{2}}{3}$. $\square$

4.D.12. 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 $\left[\frac{16}{5}, \frac{9}{5}\right]$ and $\left[-\frac{16}{5}, -\frac{9}{5}\right]$. The tangents are polars of those points. They have equations $x + y = 5$ and $x + y = -5$. $\square$

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 $x + y = \pm 5$. $\square$

4.D.13. 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 $z = 0$. $\square$

4.D.14. 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 $\frac{a^2 q^2 - b^2 p^2}{a^2 + b^2} = \frac{a^2 b^2}{a^2 + b^2}$, because P lies on hyperbola. $\square$

4.D.15. 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. $\square$

4.D.16. 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 $\left[\frac{2}{3}, -2\right]$.

(b) In this case,

$$\begin{aligned} s_1 + s_2 + 1 &= 0 \\ s_1 + s_2 + \tfrac{1}{2} &= 0. \end{aligned}$$

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 (α, β). The coordinates of the center therefore give the translation of the coordinate system to the frame in which the ellipse has its basic form.

$\square$

4.D.17. 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 $-9x - 21y - 5 = 0$. $\square$

4.D.18. 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. $\square$

4.D.19. The equation

$$x^2 + 3xy - y^2 + x + y + 1 = 0.$$

defines a conic section. Determine its center, axes, asymptotes and foci.

4.D.20. 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 homogenenous 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 $y = 1$.

$\square$

4.D.21. 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 $F(p, v) = p^T A v = 0$, 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.22. 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.23. 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 $t^T A t = 0$, which gives

$$2t_1^2 - 4t_1 t_2 + t_2^2 - 2t_1 t_3 + 6t_2 t_3 - 3t_3^2 = 0$$

The condition that P lies on the polar of T is $p^T A t = 0$, 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

$$-t_1^2 + 4t_1 t_3 - 3t_3^2 = 0$$

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

$$-(\tfrac{t_1}{t_3})^2 + 4(\tfrac{t_1}{t_3}) - 3 = 0 \quad \text{a} \quad \tfrac{t_2}{t_3} = 3(\tfrac{t_1}{t_3}) - 6,$$

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.24. 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 $20x - 21y = 0$. $\qquad\square$

4.D.25. 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 $[-2, -1]$. $\qquad\square$

Solution. Alternative. The point $P = \sqrt{5}(\cos\theta, \sin\theta)$ lies on the circle for all θ. 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 $[-2, -1]$. $\qquad\square$

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.26. 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$= (-1 : 2 : 0)$ a Q$= (-3 : 2 : 0)$.

$$(-1, 2, 0) \begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -12x + 10y - 2 = 0$$

a

$$(-3, 2, 0) \begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -20x + 18y - 2 = 0$$

$\qquad\square$

There are further exercises on conic sections on the page .

***4.D.27.* 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.

Exercise solution

4.A.9. 2, 3, 4, 6, 7, 8. Find planes the positions of which correspond to each of those numbers.

4.B.12. For the normal vector (a, b, c) of such planes $ax + by + cz = d$, $a + b = 0$ since (a, b, c) must be orthogonal to the direction of p. $a = d$ since the plane contains $(1, 0, 0)$. So the plane is $ax - ay + cz = a$. If $a = 0$, then the plane is $z = 0$, The angle condition requires $\cos 60° = \frac{1}{2} = \frac{a+b+c}{\sqrt{a^2+b^2+c^2}\sqrt{3}}$

and by choosing $a = -b = 1$ (vector $(0, 0, 1)$ does not satisfy the conditions, so by certain multiplication we can get $a = -b = 1$) we then get, using the angle condition, $\left|\frac{c}{\sqrt{3}\sqrt{2+c^2}}\right| = \frac{1}{2}$, alltogether, the sought equations are $x - y \pm \sqrt{6} - 1 = 0$.

4.B.17. $(-1, 3, 2)$.

4.D.1. Line $(2t, t, 7t) + [-5, 0, -9]$.

4.D.2. $[3, 2, 1][8/3, 8/3, 2/3]$.

4.D.3. The transversal $[1, 1, 1][-3, 1, -1]$ is of length $\sqrt{20}$, so the stick is not long enough.

4.D.4. $\frac{2\sqrt{6}}{9}$

4.D.5. $\frac{\sqrt{3}}{6}$.

4.D.6. $\frac{\sqrt{3}}{\sqrt{11}}$

Establishing the ZOO

which functions do we need for our models?
– a thorough menagerie

A. Polynomial interpolation

Let us start with some examples which will hopefully make us more comfortable with polynomials.

5.A.1. Determine the sum of coefficients of the polynomial $(1 - 2x + 3x^2 - x^3)^r$, where r is your age in years.

Solution. The sum of coefficients of a polynomial is equal it's value in 1. Therefore the sum is $(1 - 2 + 3 - 1)^r = 1^r = 1$. $\square$

5.A.2. Determine the coefficient by x^{120} of the polynomial

$$P(x) = (1 - x + x^2 - x^3 + \cdots - x^{49})(1 + x + x^2 + \cdots x^{101}).$$

Solution.

$$P(x) = \frac{1 - x^{50}}{1 + x} \frac{1 - x^{102}}{1 - x} = (1 - x^{50})\frac{1 - x^{102}}{1 - x^2}$$
$$= (1 - x^{50})(1 + x^2 + x^4 + \cdots + x^{100})$$
$$= 1 + x^2 + \cdots + x^{48} - x^{102} - \cdots - x^{150}.$$

The coefficient by x^{120} is -1. $\square$

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 $\mathbb{N} \to \mathbb{K}$ or $\mathbb{Z} \to \mathbb{K}$, 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 $\mathbb{R} \to \mathbb{R}$ (*real-valued functions of one real variable*), or $\mathbb{R} \to \mathbb{C}$ (*complex-valued functions of one real variable*), or sometimes more generally the *vector-valued functions of one real variable* $\mathbb{R} \to V$. The results can usually be

5.A.3. Prove that any real solution x_0 of the equation $x^3 + px + q = 0$ $(p, q \in \mathbb{R})$ safisfies the inequality $4qx_0 \leq p$.

Solution. Note that x_0 is the solution of the quadratic equation $x_0 x^2 + px + q = 0$, therefore its discriminant $p^2 - 4x_0 p$ is non-negative. $\square$

5.A.4. Let $P(x)$ be a polynomial of degree at most n, $n \geq 1$, such that
$$P(k) = \frac{n+1-k}{k+1}$$
for $k = 0, 1, \ldots, n$. Find $P(n+1)$.

Solution. Let $Q(x) = (x+1)P(x) - (n+1-x)$. Note that $Q(x)$ has degree $n + 1$ and the condition from the problem statement now sais that $(n+1)$ numbers $0, 1, \ldots, n$ are roots of the polynomial, that is $Q(x) = K \cdot x \cdot (x-1) \cdots (x-n)$. Now we use the two expressions of $Q(x)$ to determine K. On one hand $Q(-1) = -(n+2)$, but $Q(-1) = K \cdot (-1)^{n+1} \cdot (n+1)!$ as well. Thus

$$K = \frac{(-1)^n (n+2)}{(n+1)!}$$

that is $Q(n+1) = (-1)^n (n+2)$. On the other hand from our definition of $Q(x)$ we get $Q(n+1) = (n+2)P(n+1)$. All together $P(n+1) = (-1)^n$. $\square$

5.A.5. Let $P(x)$ be a polynomial with real non-negative coefficients. Prove, that if $P(\frac{1}{x})P(x) \geq 1$ for $x = 1$ than the same inequality holds for every positive x.

Solution. Let $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots a_1 x + a_0$. From the problem statement we have $P(1)^2 \geq 1$. Further

$$P(x)P\left(\frac{1}{x}\right) = (a_n x^n + a_{n-1} x^{n-1} + \cdots a_1 x + a_0)$$

$$\cdot (a_n x^{-n} + a_{n-1} x^{-(n-1)} + \cdots a_1 x^{-1} + a_0)$$

$$= \sum_{i=0}^{n} a_i^2 + \sum_{i<j} a_i a_j (x^{j-i} + x^{i-j})$$

$$\geq \sum_{i=0}^{n} a_i^2 + 2 \sum_{i<j} a_i a_j = P(1)^2 \geq 1$$

where we have used the well known inequality $x + \frac{1}{x} \geq 2$ which holds for any positive real number x (equivalent to $(\sqrt{x} - \frac{1}{\sqrt{x}})^2 \geq 0$). $\square$

Now we will try to approximate functions by polynomials. Suppose we have incomplete information about an unknown function, namely the values it takes at several points, or the values of its first or second derivatives at those points

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 $\mathbb{Q}$.

A *polynomial* over a ring of scalars $\mathbb{K}$ is a mapping $f : \mathbb{K} \to \mathbb{K}$ given by the expression

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where a_i, $i = 0, \ldots, n$, 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 $a_0, a_1, \ldots$ 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:

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 12.2.6. Write $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ for the polynomial of degree n, and $g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$, with $a_n \neq 0$ and $b_m \neq 0$.

as well. We will try to find a polynomial (of the least degree possible) satisfying these dependencies.

5.A.6. Find a polynomial P satisfying the following conditions:
$$P(2) = 1, P(3) = 0,$$
$$P(4) = -1, P(5) = 6.$$

Solution. First, let us solve this task by creating a system of four linear equations in four variables. Suppose the polynomial is of the form $a_3 x^3 + a_2 x^2 + a_1 x_1 + a_0$. We know there is exactly one polynomial of degree less than four and satisfying the given conditions.

$$a_0 + 2a_1 + 4a_2 + 8a_3 = 1$$
$$a_0 + 3a_1 + 9a_2 + 27a_3 = 0$$
$$a_0 + 4a_1 + 16a_2 + 64a_3 = -1$$
$$a_0 + 5a_1 + 25a_2 + 125a_3 = 6.$$

Each equation arose from one of the given conditions.

Another option is to construct the required polynomial from the fundamental Lagrange polynomials.

(see 5.1.4):

$$P(x) = 1 \cdot \frac{(x-3)(x-4)(x-5)}{(2-3)(2-4)(2-5)} + 0 \cdot (\dots)$$
$$+ (-1) \cdot \frac{(x-2)(x-3)(x-5)}{(4-2)(4-3)(4-5)}$$
$$+ 6 \cdot \frac{(x-2)(x-3)(x-4)}{(5-2)(5-3)(5-4)}$$
$$= \frac{4}{3}z^3 - 12z^2 + \frac{101}{3}z - 29.$$

The coefficients of the polynomial form, of course, the solution of the aforementioned system of linear equations. □

The methods from the previous example can be applied to complex valued polynomials as well:

5.A.7. Find a polynomial P satisfying the following conditions:
$$P(1+i) = i, P(2) = 1, P(3) = -i.$$

○

5.A.8. For pairwise distinct points $x_0, \dots, x_n \in \mathbb{R}$, consider the elementary Lagrange polynomials (5.1.4):
$$l_i(x) = \frac{(x-x_0)\cdots(x-x_{i-1}(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)},$$
$x \in \mathbb{R}, i = 0, \dots, n.$
Prove that
$$\sum_{i=0}^{n} l_i(x) = 1 \quad \text{for all } x \in \mathbb{R}.$$

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. □

If $f(b)$ equals zero for some element $b \in \mathbb{K}$, then $0 = f(b) = q(b) \cdot 0 + r$, so $r = 0$. Consequently $f(x) = (x-b)q(x)$. 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. □

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

$$f(x) = (x - x_0)(x - x_1) \dots (x - x_n).$$

This is zero at these points and only at them. However, there are other polynomials which are zero at the given points. For

Solution. Apparently,

$$\sum_{i=0}^{n} l_i(x_0) = 1 + 0 + \cdots + 0 = 1,$$

$$\sum_{i=0}^{n} l_i(x_1) = 0 + 1 + \cdots + 0 = 1,$$

$$\vdots$$

$$\sum_{i=0}^{n} l_i(x_n) = 0 + 0 + \cdots + 1 = 1.$$

This means that the polynomial $\sum_{i=0}^{n} l_i(x)$ of degree not greater than n takes the value 1 at the $n+1$ points $x_0, \ldots, x_n$. However, there is exactly one such polynomial, namely the constant polynomial $y \equiv 1$. □

5.A.9. Find a polynomial P satisfying the following conditions:

$$P(1) = 0, P'(1) = 1, P(2) = 3, P'(2) = 3.$$

Solution. Once again, we will provide two methods of finding the polynomial.

The given conditions give rise to four linear equations for the coefficients of the wanted polynomial. So if we look for a polynomial of degree less than four, we get the same number of equations and unknown coefficients (let us say $P(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0$):

$$P(1) = a_3 + a_2 + a_1 + a_0 = 0,$$

$$P'(1) = 3a_3 + 2a_2 + a_1 = 1,$$

$$P(2) = 8a_3 + 4a_2 + 2a_1 + a_0 = 3,$$

$$P'(2) = 12a_3 + 4a_2 + a_1 = 3.$$

By solving this system, we obtain the polynomial $P(x) = -2x^3 + 10x^2 - 13x + 5$.

Another solution. We will use fundamental Hermite polynomials:

$$h_1^1(x) = \left(1 - \frac{2}{0 + (-1)}(x-1)\right)(2-x)^2$$
$$= (2x-1)(x-2)^2,$$

$$h_2^1(x) = (5 - 2x)(x-1)^2,$$

$$h_1^2(x) = (x-1)(x-2)^2,$$

$$h_2^2(x) = (x-2)(x-1)^2.$$

Altogether,

$$P(x) = 0 \cdot h_1^1(x) + 3 \cdot h_2^1(x) + 1 \cdot h_1^2(x) + 3 \cdot h_2^2(x)$$

$$= -2x^3 + 10x^2 - 13x + 5.$$
□

instance the zero polynomial, which is 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 $x_0, \ldots, x_n \in \mathbb{K}$ and given values $y_0, \ldots, y_n \in \mathbb{K}$ is either the zero polynomial, or a polynomial of degree at most n such that $f(x_i) = y_i$ for all $i = 0, 1, \ldots, n$.

Theorem. *For every set of $n+1$ (pairwise distinct) points $x_0, \ldots, x_n \in \mathbb{K}$ and given values $y_0, \ldots, y_n \in \mathbb{K}$, there is exactly one interpolation polynomial f.*
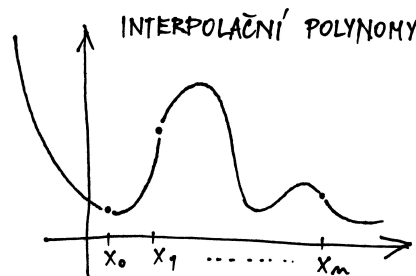
PROOF. If f and g are interpolation polynomials with the same defining values, then their difference is a polynomial of degree n which has at least $n+1$ roots, and thus $f - g = 0$. This proves uniqueness.

It remains to prove the existence. Label the coefficients of the polynomial f of degree n:

$$f = a_n x^n + \cdots + a_1 x + a_0.$$

Substituting the desired values leads to a system of $n+1$ equations for the same number of unknown coefficients a_i

$$a_0 + x_0 a_1 + \cdots + (x_0)^n a_n = y_0$$

$$\vdots$$

$$a_0 + x_n a_1 + \cdots + (x_n)^n a_n = y_n.$$

The existence of a solution of this system is easily shown by constructing the polynomial by using the Lagrange polynomials for the given points $x_0, \ldots, x_n$. (See below).

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.B.7 on page 89.

Since it is verified that for zero right-hand sides, there is exactly one solution, we know that this determinant must be non-zero.

Since polynomials are equal as mappings if and only if they are equal as sequences of coefficients, the theorem is proved. □

5.A.10. Using Lagrange interpolation, approximate $\cos^2 1$. Use the values the function takes at the points $\frac{\pi}{4}$, $\frac{\pi}{3}$, and $\frac{\pi}{2}$.

Solution. First, we determine the mentioned values: $\cos^2(\frac{\pi}{4}) = 1/2$, $\cos^2(\frac{\pi}{3}) = 1/4$, $\cos^2(\frac{\pi}{2}) = 0$. Then, we determine the elementary Lagrange polynomials, calculating their values at the given point.

$$l_0(1) = \frac{(1 - \frac{\pi}{3})(1 - \frac{\pi}{2})}{(\frac{\pi}{4} - \frac{\pi}{3})(\frac{\pi}{4} - \frac{\pi}{2})} = 8\frac{(\pi - 3)(\pi - 2)}{\pi^2},$$

$$l_1(1) = \frac{(1 - \frac{\pi}{4})(1 - \frac{\pi}{2})}{(\frac{\pi}{3} - \frac{\pi}{4})(\frac{\pi}{3} - \frac{\pi}{2})} = -9\frac{(\pi - 4)(\pi - 2)}{\pi^2},$$

$$l_2(1) = \frac{(1 - \frac{\pi}{4})(1 - \frac{\pi}{3})}{(\frac{\pi}{2} - \frac{\pi}{4})(\frac{\pi}{2} - \frac{\pi}{3})} = 2\frac{(\pi - 4)(\pi - 3)}{\pi^2}.$$

Altogether,

$$P(1) = \frac{1}{2} \cdot 8\frac{(\pi - 3)(\pi - 2)}{\pi^2} - \frac{1}{4} \cdot 9\frac{(\pi - 4)(\pi - 2)}{\pi^2} + 0 =$$

$$= \frac{(7\pi - 12)(\pi - 2)}{4\pi^2} \doteq 0.288913.$$

We may notice we did not need to calculate the third elementary polynomial. The actual value is $\cos^2 1 \doteq 0.291927$.
□

5.A.11. Joe needs to calculate values of the sine function with a calculator capable of basic arithmetic operations only. As he remembers the sine's values at the points 0, $\frac{\pi}{6}$, $\frac{\pi}{4}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$ and knows that π, $\sqrt{2}$, and $\sqrt{3}$ are approximately 3.1416, 1.4142, and 1.7321, respectively, he decided to use interpolation. Help him build an approximate formula, using all of the given values.

Solution. We will construct the elementary Lagrange polynomials:

$$l_0(x) = \frac{(x - \frac{\pi}{6})(x - \frac{\pi}{4})(x - \frac{\pi}{3})(x - \frac{\pi}{2})}{(0 - \frac{\pi}{6})(0 - \frac{\pi}{4})(0 - \frac{\pi}{3})(0 - \frac{\pi}{2})}$$
$$\doteq 1.4783x^4 - 5.8052x^3 + 8.1057x^2 - 4.7746x + 1,$$

$$l_1(x) = \frac{(x - 0)(x - \frac{\pi}{4})(x - \frac{\pi}{3})(x - \frac{\pi}{2})}{(\frac{\pi}{6} - 0)(\frac{\pi}{6} - \frac{\pi}{4})(\frac{\pi}{6} - \frac{\pi}{3})(\frac{\pi}{6} - \frac{\pi}{2})}$$
$$\doteq -13.3046x^4 + 45.2808x^3 - 49.2419x^2 + 17.1887x,$$

$$l_2(x) = \frac{(x - 0)(x - \frac{\pi}{6})(x - \frac{\pi}{3})(x - \frac{\pi}{2})}{(\frac{\pi}{4} - 0)(\frac{\pi}{4} - \frac{\pi}{6})(\frac{\pi}{4} - \frac{\pi}{3})(\frac{\pi}{4} - \frac{\pi}{2})}$$
$$\doteq 23.6526x^4 - 74.3070x^3 + 71.3298x^2 - 20.3718x,$$

$$l_3(x) = \frac{(x - 0)(x - \frac{\pi}{6})(x - \frac{\pi}{4})(x - \frac{\pi}{2})}{(\frac{\pi}{3} - 0)(\frac{\pi}{3} - \frac{\pi}{6})(\frac{\pi}{3} - \frac{\pi}{4})(\frac{\pi}{3} - \frac{\pi}{2})}$$
$$\doteq -13.3046x^4 + 38.3146x^3 - 32.8279x^2 + 8.5943x,$$

$$l_4(x) = \frac{(x - 0)(x - \frac{\pi}{6})(x - \frac{\pi}{4})(x - \frac{\pi}{3})}{(\frac{\pi}{2} - 0)(\frac{\pi}{2} - \frac{\pi}{6})(\frac{\pi}{2} - \frac{\pi}{4})(\frac{\pi}{2} - \frac{\pi}{3})}$$
$$\doteq 1.4783x^4 - 3.4831x^3 + 2.6343x^2 - 0.6366x.$$

5.1.4. Applications of interpolations. At first sight, it may seem that real or rational polynomials, that is, polynomial functions $\mathbb{R} \to \mathbb{R}$ or $\mathbb{Q} \to \mathbb{Q}$, 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 polynomial related to fixed points $x_0, \ldots, x_n$:

LAGRANGE[1] INTERPOLATION POLYNOMIALS

The *Lagrange interpolation polynomial* is expressed in terms of the *elementary Lagrange polynomials* ℓ_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 $(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)$. 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

$$f(x) = y_0\ell_0(x) + y_1\ell_1(x) + \cdots + y_n\ell_n(x).$$

dat nekam poznamku o pocitani pres vzorecek s derivaci

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 ℓ_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.

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).

---

[1]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.

Then, the value of the interpolation polynomial $P(x)$ is

$$0 \cdot l_0(x) + \frac{1}{2}l_1(x) + \frac{\sqrt{2}}{2}l_2(x) + \frac{\sqrt{3}}{2}l_3(x) + l_4(x)$$
$$\doteq 0.0288x^4 - 0.2043x^3 + 0.0214x^2 + 0.9956x.$$

$\square$

Additional questions: Can Joe use this formula to calculate the sine's values at the interval $\left[\frac{\pi}{2}, \pi\right]$? If not, what should he do?

What would the approximate formulae look like if he used not all five knots, but only the three nearest ones for each point?

5.A.12. The day after, Joe needed to calculate the binary logarithm of 25. (Actually, he needed the natural logarithm of 25, but since he knows that $\ln 2$ is approximately 0.6931, the binary one will do.) So he took the points 16 and 32 (with values 4 and 5, respectively) and constructed the interpolation polynomial (line). $P(x) = \frac{1}{16}x + 3$, hence $P(25) = \frac{73}{16} = 4.5625$. Then, he added the point 8 (with value 3) in order to arrive at a more accurate result. In this case, the interpolation polynomial equals $P(x) = -\frac{1}{384}x^2 + \frac{3}{16}x + \frac{5}{3}$, which gives $P(25) \doteq 4.7266$. Joe wanted to obtain an even more accurate number, so he added two more points, namely 2 and 4 (with values 1 and 2, respectively). How shocked he was when he got the result $P(25) \doteq 5.892$, which is apparently wrong as the binary logarithm is an increasing function. Can you explain the origin of this error?

Solution. Joe asked Google and learned that the interpolation error can be expressed as

$$f(x) - P_n(x) = \frac{(x - x_0)(x - x_1)\dots(x - x_n)}{(n+1)!} f^{(n+1)}(\xi),$$

where the point ξ is not known, but lies in the interval given by the least and greatest knots. The term in the fraction's numerator causes the accuracy to deteriorate by adding farther knots. $\square$

5.A.13. A week later, Joe needed to approximate $\sqrt{7}$. He got the idea of reversing the problem and using the inverse interpolation, ie. to interchange the roles of arguments (function inputs) and values (function outputs) and to approximate the value of an appropriate function at the point 0. Describe his procedure.

Solution. The function $x^2 - 7$ takes 0 at $\sqrt{7}$. Joe took the points $x_0 = 2$, $x_1 = 2.5$, and $x_2 = 3$, with the function values -3, -0.75, and 2, respectively. Then he interchanged

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.
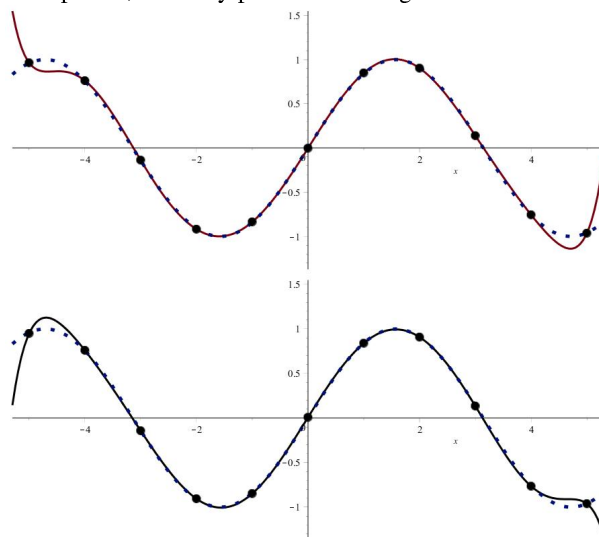
There is a rich theory about the interpolation polynomials. If interested, consult the special literature.

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 2.2.9). This exhibits the desired determinant as the polynomial

(1)
$$V(x_0, \dots, x_n) = (x_n)^n V(x_0, \dots, x_{n-1}) -$$
$$(x_n)^{n-1} V(x_0, \dots, x_{n-2}, x_n) + \cdots.$$

their roles, thus obtaining the elementary Lagrange polynomials

$$l_0(x) = \frac{(x+0.75)(x-2)}{(-3+0.75)(-3-2)} = \frac{4}{45}x^2 - \frac{1}{9}x - \frac{2}{15},$$

$$l_1(x) = -\frac{16}{99}x^2 - \frac{16}{99}x + \frac{32}{33},$$

$$l_2(x) = \frac{6}{55}x^2 + \frac{3}{11}x + \frac{9}{55}.$$

For $\sqrt{7}$, he got the approximate value $2 \cdot l_0(0) + 2.5 \cdot l_1(0) + 3 \cdot l_2(0) = \frac{437}{165} \doteq 2.6485$.

Additional questions: Joe made a mistake while constructing one of the elementary polynomials, try to find it. Does this mistake affect the resulting value?

How could we make use of the value of the derivative at the point 2.5? $\qquad\square$

Finding a spline through given data is a tedious task for the hand computation (if we are given n ($n \geq 2$) points and values in them, then we need to solve $4n-4$ linear equations. The matrix of this system is special though (see 5.1.9) and there are algorithms to transform the task to solve actually only n linear equations with n unknowns.

We show some "ad hoc" approach which can even more simplify the problem in a special situation.

5.A.14. Find a natural spline S which satisfies
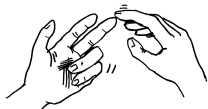
$$S(-1) = 0,\ S(0) = 1,\ S(1) = 0.$$

Solution. The wanted spline consists of two cubic polynomials, let us denote them S_1 for the interval $[-1, 0]$ and S_2 for the interval $[0, 1]$. The word "natural" requires that the second derivatives of $S_1 = ax^3 + bx^2 + cx + d$ and $S_2 = ex^3 + fx^2 + gx + h$ be zero at the points -1 and 1, respectively. Applying the definition of spline only, we get eight linear equations. We can reduce the system 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. But the 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 we have $S_1(x) = ax^3 + bx^2 + 1$ and $S_2(x) = -ax^3 + bx^2 + 1$, Confronting these forms with the conditions

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

$$V(x_0, \ldots, x_n) = c \cdot (x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}).$$

Comparing the coefficients at the highest power in (1) with this expression yields

$$c = V(x_0, \ldots, x_{n-1}),$$

which completes the proof. $\qquad\square$

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

$$(1) \qquad \frac{f(x + \delta x) - f(x)}{\delta x},$$

where δx is a small value in $\mathbb{K}$ expressing the increment of the argument x. Since we can calculate (over an arbitrary ring)

$$(x + \delta x)^k = x^k + \cdots + \binom{k}{l}x^l(\delta x)^{k-l} + \cdots + (\delta x)^k,$$

we get for the polynomial $f(x) = a_n x^n + \cdots + a_0$, the above quotient (1) in the form

$$\frac{f(x+\delta x)-f(x)}{\delta x} = a_n\frac{nx^{n-1}\delta x+\cdots+(\delta x)^k}{\delta x}+\cdots+a_1\frac{\delta x}{\delta x}$$

$$= na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1 + \delta x(\ldots)$$

where the expression in parentheses in the end of the expression is polynomially dependent on δx. Clearly, for values δx very close to zero, there is a value arbitrarily close to the following expression:

<div style="text-align:center">DERIVATIVES OF POLYNOMIALS</div>

The derivative of the polynomial $f(x) = a_n x^n + \cdots + a_0$ with respect to the variable x is the polynomial

$$f'(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1.$$

obrazek, treba uplne explicitne pro kvadraticky pripad

$S_1(-1) = 0$ and $S_1"(-1) = 0$ yields the following system of only two linear equations in a and b.

$$-a + b + 1 = 0,$$
$$-6a + 2b = 0.$$

Having solved that, we get $S_1(x) = -\frac{1}{2}x^3 - \frac{3}{2}x^2 + 1$, $S_2(x) = \frac{1}{2}x^3 - \frac{3}{2}x^2 + 1$. Altogether,

$$S(x) = \begin{cases} -\frac{1}{2}x^3 - \frac{3}{2}x^2 + 1 & \text{pro } x \in [-1, 0], \\ \frac{1}{2}x^3 - \frac{3}{2}x^2 + 1 & \text{pro } x \in [0, 1]. \end{cases}$$

$\square$

You can use the same trick to solve the following problem.

5.A.15. Find a (cubic) spline S which satisfies

$S(-1) = 0$, $S(0) = 1$, $S(1) = 0$, $S'(-1) = -1$, $S'(1) = 1$.

$\bigcirc$

5.A.16. Find a polynomial of degree two or less such that its values at the points

$$x_0 = -1, \quad x_1 = 1, \quad x_2 = 2$$

are

$$y_0 = 1, \quad y_1 = -3, \quad y_2 = 4,$$

respectively.

$\bigcirc$

5.A.17. Construct the Lagrange interpolation polynomial for

x_i	-2	-1	1	2
y_i	1	-1	-1	1

Then find any polynomial of degree greater than three which satisfies the conditions in the table.

$\bigcirc$

5.A.18. Find a polynomial $p(x) = ax^3 + bx^2 + cx + d$ which satisfies $p(0) = 1$, $p(1) = 0$, $p(2) = 1$, $p(3) = 10$.

$\bigcirc$

5.A.19. Construct a polynomial p of degree three or less which satisfies $p(0) = 2$, $p(1) = 3$, $p(2) = 12$, $p(5) = 147$.

$\bigcirc$

5.A.20. Let the values $y_0, \ldots, y_n \in \mathbb{R}$ at pairwise distinct points $x_0, \ldots, x_n \in \mathbb{R}$, respectively, be given. How many polynomials of degree exactly $n+1$ and taking the given values at the given points are there?

$\bigcirc$

5.A.21. Determine the Hermite interpolation polynomials P, Q satisfying

$P(-1) = -11$, $P(1) = 1$, $P'(-1) = 12$, $P'(1) = 4$;
$Q(-1) = -9$, $Q(1) = -1$, $Q'(-1) = 10$, $Q'(1) = 2$. $\bigcirc$

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 δx 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 :

$$\big(f(x) \cdot g(x)\big)' = f'(x) \cdot g(x) + f(x) \cdot g'(x).$$

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 12.2.7.

5.1.7. Hermite's interpolation problem. Consider $m + 1$ pairwise distinct real numbers $x_0, \ldots, x_m$, 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. A polynomial f is wanted which will satisfy these conditions on the values and derivatives.

picture missing!

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 $f(x) = a_n x^n + \cdots + a_0$:

$$a_0 + x_0 a_1 + \cdots + (x_0)^n a_n = y_0$$

$$\vdots$$

$$a_0 + x_m a_1 + \cdots + (x_m)^n a_n = y_m$$

$$a_1 + 2x_0 a_2 + \cdots + n(x_0)^{n-1} a_n = y_0'$$

$$\vdots$$

$$a_1 + 2x_m a_2 + \cdots + n(x_m)^{n-1} a_n = y_m'.$$

5.A.22. Replace the function f with a Hermite polynomial, knowing following values of f:

x_i	-1	1	2
$f(x_i)$	4	-4	-8
$f'(x_i)$	8	-8	11

In the following excercises let $y_i := f(x_i)$.

5.A.23. Without calculation, determine the Hermite interpolation polynomial if the following is given:

$$x_0 = 0,\ x_1 = 2,\ x_2 = 1,$$
$$y_0 = 0,\ y_1 = 4,\ y_2 = 1,$$
$$y_0' = 0,\ y_1' = 4,\ y_2' = 2.$$

5.A.24. Find a polynomial of degree three or less taking the value $y = 4$ at the point $x = 1$ and $y = 9$ at $x = 2$, having its derivative equal to -2 at $x = 0$ and to 1 at $x = 1$. Then find a polynomial of degree three or less taking the value $y = 6$ at both the points $x = 1$ and $x = -1$ and having its derivative equal to 2 at both these points.

5.A.25. How many polynomials satisfying the following conditions are there? The degree is four or less, the values at $x_0 = 5$ and $x_1 = 55$ are $y_0 = 55$ and $y_1 = 5$, respectively, and both the first and second derivatives at the point x_0 are zero.

5.A.26. Find any polynomial P satisfying

$$P(0) = 6, \quad P(1) = 4, \quad P(2) = 4, \quad P'(2) = 1.$$

5.A.27. Construct the natural cubic interpolation spline for the values $y_0 = 1$, $y_1 = 0$, $y_2 = 1$ at the points $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, respectively.

5.A.28. Construct the natural cubic interpolation spline for the function

$$f(x) = |x|, \quad x \in [-1, 1],$$

selecting the points $x_0 = -1$, $x_1 = 0$, $x_2 = 1$.

5.A.29. 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.A.30. Without calculation, construct the natural cubic interpolation spline for the points $x_0 = -1$, $x_1 = 0$ a $x_2 = 2$ and the value $y_0 = y_1 = y_2 = 1$ at these points.

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 th 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*[2] of the second type.

Finally we look for for a polynomial which has zero derivatives at all points x_i with the same values as ℓ_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 the value is 1. The derivative is

$$\left(h_i^{(1)}(x)\right)' = -a(\ell_i(x))^2 + (1 - a(x - x_i))2\ell_i(x)\ell_i'(x).$$

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:

[2]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.

5.A.31. Construct the complete (i. e., the derivatives at the marginal points are given) cubic interpolation spline for the points $x_0 = -3$, $x_1 = -2$, $x_2 = -1$ and the values $y_0 = 0$, $y_1 = 1$, $y_2 = 2$, $y_0' = 1$, $y_2' = 1$. ○

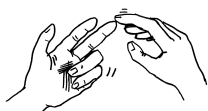**5.A.32.** Construct the natural cubic interpolation spline for the function

$$y = \frac{1}{1+x^2},$$

selecting the points $x_0 = 0$, $x_1 = 1$, $x_2 = 3$. ○

More problems concerning polynomial interpolation can be found at .

B. Topology of real numbers and their subsets

5.B.1. Find limit, isolated, boundary, and interior points of the sets $\mathbb{N}$, $\mathbb{Q}$, $X = \{x \in \mathbb{R}; 0 \le x < 1\}$ as subsets of $\mathbb{R}$.

Solution. The set $\mathbb{N}$. For any $n \in \mathbb{N}$, we have that

$$\mathcal{O}_1(n) \cap \mathbb{N} = (n-1, n+1) \cap \mathbb{N} = \{n\}.$$

Hence, there is a neighborhood of $n \in \mathbb{N}$ in $\mathbb{R}$ which contains only one natural number (the number n), therefore every point $n \in \mathbb{N}$ is isolated. There are thus no interior points (an isolated point cannot be interior). A point $a \in \mathbb{R}$ is a limit point of A if and only if every neighborhood of a contains infinitely many points of A. However, the set

$$\mathcal{O}_1(a) \cap \mathbb{N} = (a-1, a+1) \cap \mathbb{N}, \quad \text{where } a \in \mathbb{R},$$

is finite, hence $\mathbb{N}$ has no limit points. By finiteness of this set, we have that

$$\delta_b := \inf_{n \in \mathbb{N}} |b-n| = \inf_{n \in \mathcal{O}_1(b) \cap \mathbb{N}} |b-n| > 0 \quad \text{for } b \in \mathbb{R} \smallsetminus \mathbb{N}.$$

Therefore, $\mathcal{O}_{\delta_b}(b) \cap \mathbb{N} = \emptyset$, so no $b \in \mathbb{R} \smallsetminus \mathbb{N}$ is a boundary point of $\mathbb{N}$. We also know that every point which is not an interior point of a given set is necessarily its boundary point. The set of $\mathbb{N}$'s boundary points thus contains $\mathbb{N}$, and so it equals $\mathbb{N}$.

The set $\mathbb{Q}$. The rational numbers are a dense subset of the real numbers. This means that for every real number, there is a sequence of rational numbers converging to it. (We can, for instance, imagine the decimal representation of a real number and the corresponding sequence whose i-th term will be the representation truncated to the first i decimal digits. Furthermore, we can suppose that the terms of this sequence are pairwise distinct, for example by deliberately changing the

---

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

$$f(x) = \sum_{i=0}^{m} \left(y_i h_i^1(x) + y_i' h_i^2(x)\right).$$

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

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The multiplication $A \cdot (y_0, y_1, y_0', y_1')^T$ gives the vector $(a_3, a_2, a_1, a_0)^T$ of coefficients of the polynomial f, i.e.

$$f(x) = (2y_0 - 2y_1 + y_0' + y_1')x^3$$
$$+ (-3y_0 + 3y_1 - 2y_0' - y_1')x^2 + y_0'x + y_0.$$

5.1.9. 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. We will not consider the details here. 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.

last digit, or by taking the representation with recurring nines rather than zeros, ie. $0.999\ldots$ for the integer 1 and so on). The set of $\mathbb{Q}$'s limit points is thus the whole $\mathbb{R}$ and every point $x \in \mathbb{R} \setminus \mathbb{Q}$ is a boundary point. Especially, we get that any δ-neighborhood

$$\mathcal{O}_\delta\left(\frac{p}{q}\right) = \left(\frac{p}{q} - \delta, \frac{p}{q} + \delta\right), \quad \text{where } p, q \in \mathbb{Z}, q \neq 0,$$

of a rational number p/q contains infinitely many rational numbers, hence there are no isolated points. The number $\sqrt{2}/10^n$ is rational for no $n \in \mathbb{N}$. Supposing the contrary (again, $p, q \in \mathbb{Z}, q \neq 0$)

$$\frac{\sqrt{2}}{10^n} = \frac{p}{q}, \quad \text{ie.} \quad \sqrt{2} = \frac{10^n p}{q},$$

we arrive at an immediate contradiction as we know that the number $\sqrt{2}$ is not rational. Every neighborhood of a rational number p/q thus contains infinitely many real numbers $p/q + \sqrt{2}/10^n$ ($n \in \mathbb{N}$) which are not rational ($\mathbb{Q}$, as a field, is closed under subtraction). Therefore, every point $p/q \in \mathbb{Q}$ is boundary as well, and there are no interior points of the set $\mathbb{Q}$.

The set $X = [0, 1)$. 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 δ_a be the less one of the two positive numbers a, $1 - a$. Considering

$$\mathcal{O}_{\delta_a}(a) = (a - \delta_a, a + \delta_a) \subseteq (0, 1), \quad a \in (0, 1),$$

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

$$\mathcal{O}_\delta(0) \cap [0, 1) = (-\delta, \delta) \cap [0, 1) = [0, \delta),$$
$$\mathcal{O}_\delta(1) \cap [0, 1) = (1 - \delta, 1 + \delta) \cap [0, 1) = (1 - \delta, 1),$$

so every δ-neighborhood of the point 0 contains some points of the interval $[0, 1)$ and some points of the interval $(-\delta, 0)$, and every δ-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 set of X's interior points is the interval $(0, 1)$ and the set of X's boundary points is the two-element set $\{0, 1\}$, as we know that no point can be both interior and boundary and that a boundary point must be an interior or limit point. $\square$

As seen in the diagrams demonstrating the instability of interpolation by a single polynomial of sufficiently large degree, 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. Furthermore, the values of the first derivatives are usually predetermined. So instead of requiring fixed values of the first derivatives, we insist on 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 $x_0 < x_1 < \cdots < x_n$ be real values at which the required values $y_0, \ldots, y_n$ are given. A *cubic interpolation spline* for this assignment is a function $S : \mathbb{R} \to \mathbb{R}$ 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 $i = 1, \ldots, n$
- $S_i(x_{i-1}) = y_{i-1}$ and $S_i(x_i) = y_i$ for all $i = 1, \ldots n$,
- $S_i'(x_i) = S_{i+1}'(x_i)$ for all $i = 1, \ldots, n-1$,
- $S_i''(x_i) = S_{i+1}''(x_i)$ for all $i = 1, \ldots, n-1$.

The cubic spline[3] 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 $2n +$

---

[3]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.

5.B.2. Determine the suprema and infima of the following sets in $\mathbb{R}$:

$$A = (-3, 0] \cup (1, \pi) \cup \{6\}; \quad B = \left\{ \frac{(-1)^n}{n^2} \; n \in \mathbb{N} \right\};$$

$$C = (-9, 9) \cap \mathbb{Q}.$$

○

5.B.3. Find $\sup A$ and $\inf A$ for

$$A = \left\{ \frac{n + (-1)^n}{n}; \; n \in \mathbb{N} \right\} \subset \mathbb{R}.$$

○

5.B.4. The following sets are given:

$$\mathbb{N} = \{1, 2, \ldots, n, \ldots\}, \quad \mathcal{M} = \left\{ -\frac{1}{n}; \; n \in \mathbb{N} \right\},$$
$$\mathcal{J} = (0, 2] \cup [3, 5] \smallsetminus \{4\}.$$

Determine $\inf \mathbb{N}$, $\sup \mathcal{M}$, $\inf \mathcal{J}$ and $\sup \mathcal{J}$ in $\mathbb{R}$. ○

5.B.5. Find a set $M \subset \mathbb{R}$ which does not have an infimum in $\mathbb{R}$ but has a supremum there. Similarly, find a set $N \subset \mathbb{R}$ which does not have a supremum in $\mathbb{R}$ but has an infimum there. ○

5.B.6. Find a subset X of the set $\mathbb{R}$ such that $\sup X \le \inf X$. ○

5.B.7. Find sets $A, B, C \subseteq \mathbb{R}$ such that $A \cap B = \emptyset$, $A \cap C = \emptyset$, $B \cap C = \emptyset$, $\sup A = \inf B = \inf C = \sup C$. ○

C. Limits

In the subsequent exercises, we will deal with calculating limits of sequences, that is what the sequences "look like at infinity". Then, if we were to determine the n-th term of a given sequence for a very large n, the limit of the sequence (supposing it exists) can approximate it very well. We devote much space to computation of limits of sequences (and limits of functions) in this exercise column, that is why they begin earlier (and end later) than in the part concerning theory.

Let us begin with limits of sequences. The needful definitions can be found at page. .

$(n-1) + (n-1)$ 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 nonzero 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.
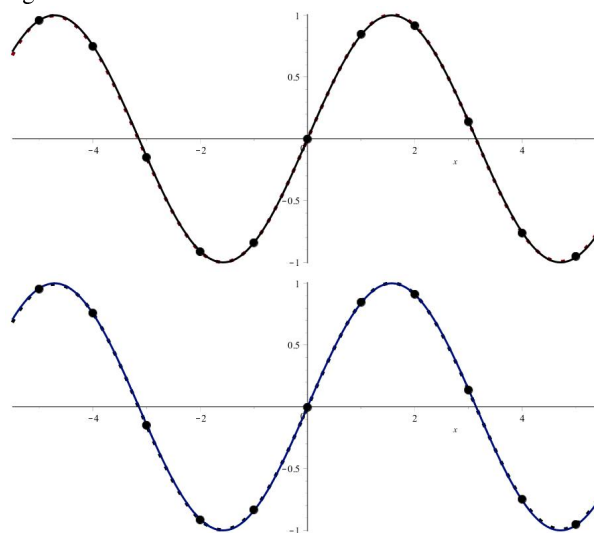
Although the diagrams look nearly identical, the data is different.

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.

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 $f : \mathbb{R} \to \mathbb{R}$ near the

5.C.1. Calculate the following limits of sequences:

i) $\lim\limits_{n\to\infty} \frac{2n^2+3n+1}{n+1}$,

ii) $\lim\limits_{n\to\infty} \frac{2n^2+3n+1}{3n^2+n+1}$,

iii) $\lim\limits_{n\to\infty} \frac{n+1}{2n^2+3n+1}$,

iv) $\lim\limits_{n\to-\infty} \frac{2^n-2^{-n}}{2^n+2^{-n}}$,

v) $\lim\limits_{n\to\infty} \frac{\sqrt{4n^2+n}}{n}$,

vi) $\lim\limits_{n\to\infty} \sqrt{4n^2+n}-2n$.

Solution.

i) $\lim\limits_{n\to\infty} \frac{2n^2+3n+1}{n+1} = \lim\limits_{n\to\infty} \frac{2n+3+\frac{1}{n}}{1+\frac{1}{n}} = \infty$.

ii) $\lim\limits_{n\to\infty} \frac{2n^2+3n+1}{3n^2+n+1} = \lim\limits_{n\to\infty} \frac{2+\frac{3}{n}+\frac{1}{n^2}}{3+\frac{1}{n}+\frac{1}{n^2}} = \frac{2}{3}$.

iii) $\lim\limits_{n\to\infty} \frac{n+1}{2n^2+3n+1} = \lim\limits_{n\to\infty} \frac{1+\frac{1}{n}}{2n+3+\frac{1}{n}} = \frac{1}{\infty} = 0$.

iv)

$$\lim_{n\to-\infty} \frac{2^n-2^{-n}}{2^n+2^{-n}} = \lim_{n\to-\infty} \frac{\frac{2^n}{2^{-n}}-1}{\frac{2^n}{2^{-n}}+1} = -1$$

v) By the squeeze theorem (5.2.12):

$\forall n \in \mathbb{N}: \frac{\sqrt{4n^2}}{n} < \frac{\sqrt{4n^2+n}}{n} < \frac{\sqrt{4n^2+n+\frac{1}{16}}}{n}$.

Then $\lim\limits_{n\to\infty} \frac{\sqrt{4n^2}}{n} = \lim\limits_{n\to\infty} \frac{2n}{n} = 2$,

$\lim\limits_{n\to\infty} \frac{\sqrt{4n^2+n+\frac{1}{16}}}{n} = \lim\limits_{n\to\infty} \frac{2n+\frac{1}{4}}{n} = 2$.

So $\lim\limits_{n\to\infty} \frac{\sqrt{4n^2+n}}{n} = 2$ as well.

vi) $\lim\limits_{n\to\infty} \sqrt{4n^2+n}-2n =$

$$\lim_{n\to\infty} \frac{(\sqrt{4n^2+n}-2n)(\sqrt{4n^2+n}+2n)}{\sqrt{4n^2+n}+2n} =$$

$$\lim_{n\to\infty} \frac{n}{\sqrt{4n^2+n}+2n} =$$

$$\lim_{n\to\infty} \frac{1}{\frac{\sqrt{4n^2+n}}{n}+2} = \frac{1}{4}.$$

$\square$

5.C.2. Let $c \in \mathbb{R}^+$ (a positive real number). We will show that $\lim\limits_{n\to\infty} \sqrt[n]{c} = 1$.

Solution. First, let us consider $c > 1$. The function $\sqrt[n]{c}$ is decreasing (in n), yet all its values are greater than 1, hence the sequence $\sqrt[n]{c}$ has a limit, and this limit is equal to the infimum of the sequence's terms. Let us suppose, for a while, that thus limit is greater than 1, that is $1 + \varepsilon$ for some $\varepsilon > 0$. Then by the definition of a limit, all the sequence's terms will eventually (from some index m on) be less than $1 + \varepsilon + \frac{\varepsilon^2}{4}$,

"jump discontinuities". For instance, the *Heaviside function*[4]

$$f(x) = \begin{cases} 0 & \text{for all } x < 0, \\ 1/2 & \text{for } x = 0, \\ 1 & \text{for all } x > 0 \end{cases}$$

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".

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

---

[4] 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?

especially $\sqrt[m]{c} < 1 + \varepsilon + \frac{eps^2}{4}$. But then we have that

$$\sqrt[2m]{c} = \sqrt{\sqrt[m]{c}} < \sqrt{1 + \varepsilon + \frac{\varepsilon^2}{4}} = 1 + \frac{\varepsilon}{2} < 1 + \varepsilon,$$

which contradicts our assumption that $1 + \varepsilon$ is the infimum of the considered sequence.

The theorem is trivial for $c = 1$, and for a number $c \in (0, 1)$ it follows from the above, if we invoke the theorem for the number $1/c$. $\square$

5.C.3. Determine

$$\lim_{n \to \infty} \sqrt[n]{n}.$$

Solution. Apparently, we have $\sqrt[n]{n} \geq 1$, $n \in \mathbb{N}$. So we can set

$$\sqrt[n]{n} = 1 + a_n \quad \text{for certain numbers} \quad a_n \geq 0, \ n \in \mathbb{N}.$$

By the binomial theorem we get that

$$n = (1 + a_n)^n = 1 + \binom{n}{1} a_n + \binom{n}{2} a_n^2 + \cdots + a_n^n,$$
$$n \geq 2 \, (n \in \mathbb{N}).$$

Hence we have the bound (all the numbers a_n are non-negative)

$$n \geq \binom{n}{2} a_n^2 = \frac{n(n-1)}{2} \, a_n^2, \quad n \geq 2 \, (n \in \mathbb{N}),$$

which leads to

$$0 \leq a_n \leq \sqrt{\frac{2}{n-1}}, \quad n \geq 2 \, (n \in \mathbb{N}).$$

By the squeeze theorem,

$$0 = \lim_{n \to \infty} 0 \leq \lim_{n \to \infty} a_n \leq \lim_{n \to \infty} \sqrt{\frac{2}{n-1}} = 0.$$

Thus we have obtained the result

$$\lim_{n \to \infty} \sqrt[n]{n} = \lim_{n \to \infty} (1 + a_n) = 1 + 0 = 1.$$

We can notice that by further application of the squeeze theorem, we get

$$1 = \lim_{n \to \infty} 1 \leq \lim_{n \to \infty} \sqrt[n]{c} \leq \lim_{n \to \infty} \sqrt[n]{n} = 1,$$

for every real number $c \geq 1$. $\square$

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.

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 necessry 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, 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 $z = \operatorname{Re} z + i \operatorname{Im} z$. 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.

5.C.4. Calculate the limit
$$\lim_{n\to\infty}\left(\sqrt{2}\cdot\sqrt[4]{2}\cdot\sqrt[8]{2}\cdots\sqrt[2^n]{2}\right).$$

Solution. To determine the limit, it is sufficient to express the terms in the form
$$2^{\frac{1}{2}}\cdot2^{\frac{1}{4}}\cdot2^{\frac{1}{8}}\cdots2^{\frac{1}{2^n}}=2^{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}}.$$

Thus we get
$$\lim_{n\to\infty}\left(\sqrt{2}\cdot\sqrt[4]{2}\cdot\sqrt[8]{2}\cdots\sqrt[2^n]{2}\right)=$$
$$\lim_{n\to\infty}2^{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}}=$$
$$2^{\lim_{n\to\infty}\left(\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}\right)}=2^{\sum\limits_{n=1}^{\infty}\frac{1}{2^n}}.$$

By the well-known formula for the sum of geometric series,
$$\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n=\frac{\frac{1}{2}}{1-\frac{1}{2}}=1,$$

whence it follows that
$$\lim_{n\to\infty}\left(\sqrt{2}\cdot\sqrt[4]{2}\cdot\sqrt[8]{2}\cdots\sqrt[2^n]{2}\right)=2^1=2.$$
$\square$

5.C.5. Determine
$$\lim_{n\to\infty}\left(\frac{1}{n^2}+\frac{2}{n^2}+\cdots+\frac{n-2}{n^2}+\frac{n-1}{n^2}\right).$$
$\bigcirc$

5.C.6. Calculate
$$\lim_{n\to\infty}\frac{\sqrt{n^3-11n^2+2}+\sqrt[5]{n^7-2n^5-n^3}-n+\sin^2 n}{2-\sqrt[3]{5n^4+2n^3+5}}.$$
$\bigcirc$

5.C.7. Determine the limit
$$\lim_{n\to\infty}\frac{n!+(n-2)!-(n-4)!}{n^{50}+n!-(n-1)!}.$$
$\bigcirc$

5.C.8. Find two sequences (let use denote their terms by x_n and y_n ($n\in\mathbb{N}$), respectively) having infinite limits and such that
$$\lim_{n\to\infty}\left(x_n+y_n\right)=1,\qquad\lim_{n\to\infty}\left(x_n\,y_n^2\right)=+\infty.$$
$\bigcirc$

5.C.9. Determine the limit points of the sequence given by
$$a_n=\frac{(-1)^n\,2n}{\sqrt{4n^2+5n+3}},\quad n\in\mathbb{N}.$$
$\bigcirc$

5.C.10. Calculate
$$\limsup_{n\to\infty}a_n\quad\text{and}\quad\liminf_{n\to\infty}a_n$$

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 $z\in\mathbb{C}$:
$$\bar{z}=\operatorname{Re}z-i\operatorname{Im}z.$$

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)
$$|z|^2=z\cdot\bar{z}.$$

The absolute value can be defined on any ordered field of scalars $\mathbb{K}$. Define the *absolute value* $|a|$ as follows:
$$|a|=\begin{cases}a&\text{if }a\geq0\\-a&\text{if }a<0.\end{cases}$$

For any numbers $a,b\in\mathbb{K}$,

(2)
$$|a+b|\leq|a|+|b|.$$

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.

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 $a_0,a_1,\ldots$, or as infinite vectors $(a_0,a_1,\ldots)$, or as $(a_i)_{i=1}^{\infty}$.

CAUCHY[5] SEQUENCES

Consider a sequence $(a_0,a_1,\ldots)$ 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*.

---

[5] 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.

if

$$a_n = \frac{n^2 + 4n - 5}{n^2 + 9} \sin^2 \frac{n\pi}{4}, \quad n \in \mathbb{N}.$$

5.C.11. Determine

$$\liminf_{n\to\infty} \left((-1)^n \left(1 + \frac{1}{n} \right)^n + \sin \frac{n\pi}{4} \right).$$

5.C.12. Now let us proceed with limits of functions. The definition can be found at page [299].

Determine

(a)

$$\lim_{x\to\pi/3} \sin x;$$

(b)

$$\lim_{x\to 2} \frac{x^2 + x - 6}{x^2 - 3x + 2};$$

(c)

$$\lim_{x\to +\infty} \left(\arccos \frac{1}{x+1} \right)^3;$$

(d)

$$\lim_{x\to -\infty} \arctan \frac{1}{x}, \quad \lim_{x\to -\infty} \arctan x^4, \quad \lim_{x\to -\infty} \arctan (\sin x).$$

Solution. (a) Let us remind that a function f is, by definition, continuous at a given point x iff the limit of f at x is equal to the function value $f(x)$. However, we know that the function $y = \sin x$ is continuous at every real number. Thus we get that

$$\lim_{x\to\pi/3} \sin x = \sin \frac{\pi}{3} = \frac{\sqrt{3}}{2}.$$

(b) The immediate substitution $x = 2$ leads to both zero numerator and zero denominator. Despite that, the problem can be solved very easily. The reduction

$$\lim_{x\to 2} \frac{x^2 + x - 6}{x^2 - 3x + 2} = \lim_{x\to 2} \frac{(x-2)(x+3)}{(x-2)(x-1)} = \lim_{x\to 2} \frac{x+3}{x-1}$$
$$= \frac{2+3}{2-1} = 5$$

leads to the correct result (thanks to continuity of the obtained at function at the point $x_0 = 2$). Let us realize that the limit of a function can be calculated from the function values in an arbitrarily small deleted neighborhood of a given point x_0 and that the limit does not depend on the function value at the point. We can thus make use of multiplying or reducing by factors which do not change the function values in an arbitrarily selected deleted neighborhood of the point x_0.

CAUCHY SEQUENCE OF COMPLEX NUMBERS
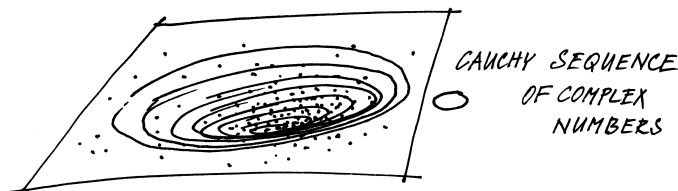
Intuitively, either all but finitely many of the sequence's terms are equal, (then $|a_i - a_j| = 0$ from some index N on), or they "approach" some particular value. This is easily imaginable in the complex plane. Choose an arbitrarily small disc (with radius equal to ε). Suppose a Cauchy sequence is given. It must be possible to put the disc into the complex plane in such a way that it covers all but a finitely many of the elements of the infinite sequence a_i. Imagine that the disc has very small radius, and contains 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 iff for any positive real number ε,

$$|a_i - a| < \varepsilon$$

for all but finitely many indices i. Notice that the set of those i for which the inequality does not hold may depend on ε). The number a is called the *limit* of the sequence $(a_i)_{i=0}^\infty$.

If a sequence $a_i \in \mathbb{K}$, $i = 0, 1, \ldots$, converges to $a \in \mathbb{K}$, then for any fixed positive ε, $|a_i - a| < \varepsilon$ for all i greater than a certain $N \in \mathbb{N}$. By the triangle inequality,

$$|a_i - a_j| = |a_i - a_N + a_N - a_j| < |a_i - a_N| + |a_N - a_j| < 2\varepsilon.$$

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 ε 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

(c) By moving the limit inwards twice, the original limit transforms to

$$\left(\arccos \left(\lim_{x \to +\infty} \frac{1}{x+1} \right) \right)^3 .$$

It can easily be shown that

$$\lim_{x \to +\infty} \frac{1}{x+1} = 0.$$

As the function $y = \arccos x$ is continuous at the point 0 and takes the value $\pi/2$ there, and the function $y = x^3$ is continuous at $\pi/2$, we get that

$$\lim_{x \to +\infty} \left(\arccos \frac{1}{x+1} \right)^3 = \left(\arccos \left(\lim_{x \to +\infty} \frac{1}{x+1} \right) \right)^3$$
$$= \left(\frac{\pi}{2} \right)^3 .$$

(d) The function $y = \arctg x$ has properties which are "useful when calculating limits" – it is continuous and injective (increasing) on the whole domain. These properties always (with no further conditions or limitations) allow to move the examined limit into the argument of such a function. Therefore, let us consider

$$\arctg \left(\lim_{x \to -\infty} \frac{1}{x} \right), \quad \arctg \left(\lim_{x \to -\infty} x^4 \right),$$
$$\arctg \left(\lim_{x \to -\infty} \sin x \right) .$$

Apparently,

$$\lim_{x \to -\infty} \frac{1}{x} = 0, \qquad \lim_{x \to -\infty} x^4 = +\infty$$

and the limit $\lim_{x \to -\infty} \sin x$ does not exist, which implies

$$\lim_{x \to -\infty} \arctg \frac{1}{x} = \arctg 0 = 0,$$
$$\lim_{x \to -\infty} \arctg x^4 = \lim_{y \to +\infty} \arctg y = \frac{\pi}{2}$$

and the last limit does not exist, either. $\qquad \square$

5.C.13. Determine the limit

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2 \sin(x^2)} .$$

Solution.

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2 \sin(x^2)} = \lim_{x \to 0} \frac{2 \sin^2 \left(\frac{x}{2} \right)}{x^2 \sin(x^2)}$$
$$= \lim_{x \to 0} \frac{\frac{1}{2} \sin^2 \left(\frac{x}{2} \right)}{\left(\frac{x}{2} \right)^2 \sin(x^2)}$$
$$= \frac{1}{2} \left(\lim_{x \to 0} \frac{\sin \left(\frac{x}{2} \right)}{\frac{x}{2}} \right)^2 \cdot \lim_{x \to 0} \frac{1}{\sin^2(x^2)}$$
$$= \frac{1}{2} \cdot \infty = \infty.$$

to B, while $a_N + \varepsilon$ does not. Hence $|a - a_N| \le \varepsilon$, and thus

$$|a - a_j| \le |a - a_N| + |a_N - a_j| \le 2\varepsilon$$

for all $j > N$. So a is the limit of the given sequence. $\qquad \square$

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 $|a_i - a_j|^2 \le |z_i - z_j|^2$ 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 $z = a + i\, b$ which is the limit of the sequence z_i. $\qquad \square$

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) 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 equivalence classes by taking 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 $a \le b$ if and only if there are representatives with $a_i \le b_i$. Finally it is necessary to show that all Cauchy sequences in $\mathbb{R}$ converge. We do not go into details here and advise the reader to return back now 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.

The previous calculation must be considered "from the back". 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. If we split the original limit into the product

$$\lim_{x\to 0}(1-\cos x) \cdot \lim_{x\to 0}\frac{1}{x^2\sin(x^2)},$$

we would get the $0 \cdot \infty$ type, which is an indeterminate form, but this tells us nothing about existence of the original limit. $\square$

5.C.14. Determine the following limits:

 i) $\lim_{x\to 2}\frac{x-2}{\sqrt{x^2-4}}$,   ii) $\lim_{x\to 0}\frac{\sin(\sin x)}{x}$,
 iii) $\lim_{x\to 0}\frac{\sin^2 x}{x}$,   iv) $\lim_{x\to 0}e^{\frac{1}{x}}$.

Solution.

 i)
$$\lim_{x\to 2}\frac{x-2}{\sqrt{x^2-4}} = \lim_{x\to 2}\frac{x-2}{\sqrt{(x-2)(x+2)}} =$$
$$\lim_{x\to 2}\frac{\sqrt{x-2}}{\sqrt{x+2}} = \frac{0}{4} = 0.$$

 ii)
$$\lim_{x\to 0}\frac{\sin(\sin x)}{x} \overset{(5.2.20)}{=} \lim_{y\to 0}\frac{\sin y}{y} = 1,$$

where we made use of the fact that $\lim_{x\to 0}\sin x = 0$.

 iii)
$$\lim_{x\to 0}\frac{\sin^2 x}{x} = \lim_{x\to 0}\sin x \cdot \lim_{x\to 0}\frac{\sin x}{x} = 0 \cdot 1 = 0,$$

again, the original limit exists because both the right-hand side limits exist and their product is well-defined.

iv) One must be cautious when calculating this limit. Both one-sided limits exist, but are different, which implies that the examined limit does not exist:

$$\lim_{x\to 0_+}e^{\frac{1}{x}} = e^{\lim_{x\to 0_+}\frac{1}{x}} = e^{\infty} = \infty,$$

$$\lim_{x\to 0_-}e^{\frac{1}{x}} = e^{\lim_{x\to 0_-}\frac{1}{x}} = e^{-\infty} = 0.$$

$\square$

In the following exercise, we will be concerned with so-called indeterminate forms. We recommend perceiving indeterminate forms as a helping concept which is only to facilitate the first approach to limit calculations because the obtained indeterminate form only means that one "has found out nothing". We know the limit of a sum 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

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"[6]. 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.)

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*

$$[a,b] = \{x \in \mathbb{R},\ a \le x \le b\}$$

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

---

[6]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.

limits, supposing the particular limits exist and do not lead to one of the following expressions $\infty - \infty, 0 \cdot \infty, 0/0, \infty/\infty$, which are called indeterminate forms. For completeness, let us add that these rules can be combined and that an expression containing an indeterminate form is itself considered an indeterminate form. For instance, the forms

$$-\infty + \infty = \infty - \infty, \quad \tfrac{-\infty}{3+\infty} = -\tfrac{\infty}{\infty},$$
$$\tfrac{0}{(-\infty)^3+\infty} = 0 \cdot (\infty - \infty)^{-1}$$

are all indeterminate, but the forms

$$-\infty - \infty, \quad \frac{0}{3+\infty}, \quad \frac{0}{(-\infty)^3 - \infty}$$

can be called "determinate" (one can immediately determine the limit – they correspond to the values $-\infty, 0, 0$, respectively).

5.C.15. Calculate

(a) $\lim_{x\to 2} \frac{x+2}{(x-2)^6}$, (b) $\lim_{x\to 2} \frac{x+2}{(x-2)^5}$,

(c) $\lim_{x\to+\infty} \left(2 + \frac{1}{x}\right)^{\frac{1}{x}}$, (d) $\lim_{x\to+\infty} x^{-x}$.

Solution. In exercise (a), the quotient of the numerator and the denominator gives us $4/0$. Expressions containing division by zero are inappropriate (later, we should be able to avoid them). Yet it leads to the result, it is not an indeterminate form. We may notice that the denominator approaches zero from the right (for $x \neq 2$ we have that $(x-2)^6 > 0$). We write this as $4/ + 0$. Thus the numerator and denominator are both positive in some deleted neighborhood of the point $x_0 = 2$ and one can say that the denominator, at the limit point, is "infinitely times less" than the numerator, that is

$$\lim_{x\to 2} \frac{x+2}{(x-2)^6} = +\infty,$$

which corresponds to setting $4/ + 0 = +\infty$ (similarly, we can set $4/ - 0 = -\infty$).

When calculating the limit of (b), one can proceed analogously. Since the numbers have the same sign, we get that

$$\lim_{x\to 2+} \frac{x+2}{(x-2)^5} = +\infty \neq -\infty = \lim_{x\to 2-} \frac{x+2}{(x-2)^5},$$

so the examined limit does not exist. We can write $4/\pm 0$ (or, more generally, $a/\pm 0, a \neq 0, a \in \mathbb{R}^*$), which is a "determinate form". When thoroughly distinguishing the symbols $+0$ and -0 from ± 0, $a/\pm 0$ for $a \neq 0$ always means the limit in question does not exist.

Exercises (c), (d). If $f(x) > 0$ for all considered $x \in \mathbb{R}$, then

$$f(x)^{g(x)} = e^{\ln\left(f(x)^{g(x)}\right)} = e^{g(x)\cdot\ln f(x)}.$$

from them may not be. On the other hand, if we restrict our attention to finite unions and consider a limit point of some 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 δ, then we call it the δ-*neighbourhood* of the point a.

Clearly, for $\mathbb{K} = \mathbb{R}$, $\mathcal{O}_\delta(a)$ is an open interval of length 2δ centered at a, while in the complex plane it is the interior of the circle with radius δ and center 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.*

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.

For real numbers, the δ–neighbourhood of a point a is the open interval of length 2δ, centered at a. In the complex plane, it is the disc of radius δ, also centered at a.

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:

295

Making use of the fact that the exponential function is continuous and injective on the whole of its domain ($\mathbb{R}$), we can replace the limit

$$\lim_{x \to x_0} f(x)^{g(x)}$$

with

$$e^{\lim_{x \to x_0} (g(x) \cdot \ln f(x))}.$$

Let us remind that either of these limits exists if and only if the other one exists. Further,

$$\lim_{x \to x_0} (g(x) \cdot \ln f(x)) = a \in \mathbb{R} \implies \lim_{x \to x_0} f(x)^{g(x)} = e^a,$$

$$\lim_{x \to x_0} (g(x) \cdot \ln f(x)) = +\infty \implies \lim_{x \to x_0} f(x)^{g(x)} = +\infty,$$

$$\lim_{x \to x_0} (g(x) \cdot \ln f(x)) = -\infty \implies \lim_{x \to x_0} f(x)^{g(x)} = 0.$$

Thus we can write

$$\lim_{x \to x_0} f(x)^{g(x)} = e^{\lim_{x \to x_0} g(x) \cdot \lim_{x \to x_0} \ln f(x)},$$

if both limits on the right-hand side exist and do not lead to the indeterminate form $0 \cdot \infty$. It is not difficult to realize that this indeterminate form can only be obtained in three cases, corresponding to the remaining indeterminate forms 0^0, ∞^0, 1^∞, when we have, respectively, that

$$\lim_{x \to x_0} f(x) = 0 \quad \text{and} \quad \lim_{x \to x_0} g(x) = 0,$$

$$\lim_{x \to x_0} f(x) = +\infty \quad \text{and} \quad \lim_{x \to x_0} g(x) = 0,$$

$$\lim_{x \to x_0} f(x) = 1 \quad \text{and} \quad \lim_{x \to x_0} g(x) = \pm\infty.$$

In other cases, knowledge (and existence) of the limits

$$\lim_{x \to x_0} f(x), \qquad \lim_{x \to x_0} g(x)$$

allows us to determine the result (having defined some more expressions)

$$\lim_{x \to x_0} f(x)^{g(x)} = \left(\lim_{x \to x_0} f(x) \right)^{\lim_{x \to x_0} g(x)}.$$

Since

$$\lim_{x \to +\infty} \left(2 + \frac{1}{x} \right) = 2, \quad \lim_{x \to +\infty} \frac{1}{x} = 0, \quad \lim_{x \to +\infty} x = +\infty,$$

we have that

$$\lim_{x \to +\infty} \left(2 + \frac{1}{x} \right)^{\frac{1}{x}} = 2^0 = 1,$$

$$\lim_{x \to +\infty} x^{-x} = \lim_{x \to +\infty} \left(\frac{1}{x} \right)^x = 0$$

or

$$\lim_{x \to +\infty} x^{-x} = \lim_{x \to +\infty} (x^x)^{-1} = 0.$$

The last result can be expressed as $0^\infty = 0$ or $\infty^\infty = \infty$, $\infty^{-1} = 0$ (let us emphasize that these are not indeterminate forms).

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 $|z| \le r$ 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 $N \cap A = \{a\}$.

OPEN COVER
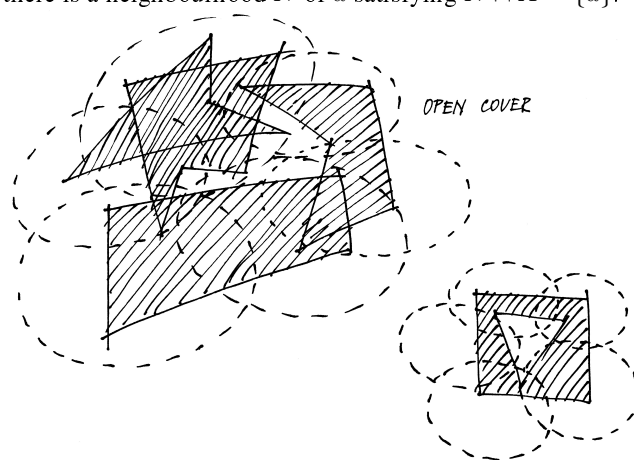
5.2.8. Theorem. *All subsets $A \subset \mathbb{K}$ of real or complex numbers satisfy:*

(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 $A \subset \mathbb{C}$ is open if and only if it is a union of countably (finite) many open circles.*

(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 it has a subsequence converging to a point in it. A,[7]*

---

[7] 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.

Although we have laid great emphasis on the reader to prefer reasoning about the limit behavior of functions to mindless labeling of the forms as determinate and indeterminate, it is, we hope, clear now why we will focus on the indeterminate ones. □

5.C.16. Calculate

$$\lim_{x\to+\infty} \frac{\sin x + \pi x^2}{2\cos x - 1 - x^2},$$

$$\lim_{x\to+\infty} \frac{3^{x+1} + x^5 - 4x}{3^x + 2^x + x^2},$$

$$\lim_{x\to+\infty} \frac{4^x - 8x^6 - 2^x - 167}{3^x - 45x - \sqrt{11}\pi^{x+12}},$$

$$\lim_{x\to+\infty} \frac{\sqrt{x} - \sin^3 x + x\,\mathrm{arctg}\,x}{\sqrt{1 + 2x + x^2}}.$$

Solution. Having reduced the first fraction by the polynomial x^2, we get

$$\lim_{x\to+\infty} \frac{\sin x + \pi x^2}{2\cos x - 1 - x^2} = \lim_{x\to+\infty} \frac{\frac{\sin x}{x^2} + \pi}{\frac{2\cos x - 1}{x^2} - 1}.$$

Boundedness of the expressions

$$|\sin x| \le 1, \quad |2\cos x - 1| \le 3 \quad \text{pro} \quad x \in \mathbb{R}$$

and $x^2 \to +\infty$ for $x \to +\infty$ give us the result

$$\lim_{x\to+\infty} \frac{\frac{\sin x}{x^2} + \pi}{\frac{2\cos x - 1}{x^2} - 1} = \frac{0 + \pi}{0 - 1} = -\pi.$$

In the last argumentation, we actually used the squeeze theorem and the notation $c/\infty = 0$ which is valid for any $c \in \mathbb{R}$ (or bounded$/\infty = 0$, where "bounded" denotes a bounded function).

This procedure can be generalized. Any limit of the form

$$\lim_{x\to x_0} \frac{f_1(x) + f_2(x) + \cdots + f_m(x)}{g_1(x) + g_2(x) + \cdots + g_n(x)},$$

where

$$\lim_{x\to x_0} \frac{f_i(x)}{f_1(x)} = 0, \quad i \in \{2,\dots,m\},$$

$$\lim_{x\to x_0} \frac{g_i(x)}{g_1(x)} = 0, \quad i \in \{2,\dots,n\},$$

satisfies

$$\lim_{x\to x_0} \frac{f_1(x) + f_2(x) + \cdots + f_m(x)}{g_1(x) + g_2(x) + \cdots + g_n(x)} = \lim_{x\to x_0} \frac{f_1(x)}{g_1(x)},$$

supposing the limit on the right-hand side exists. It is advantageous to realize (the third limit can be determined, for example, by l'Hospital's rule, with which we will make ourselves familiar later)

$$\lim_{x\to+\infty} \frac{c}{x^\alpha} = 0, \quad \lim_{x\to+\infty} \frac{x^\alpha}{x^\beta} = 0, \quad \lim_{x\to+\infty} \frac{x^\beta}{a^x} = 0,$$

(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 circles 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 circle $\mathcal{O}_\delta(z)$ contained in A, with some $\delta > 0$, and let δ_z be the supremum of the values of such δ. 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 circles $\mathcal{O}_{\delta_w}(w)$ for all $w \in A$ with rational real and imaginary coordinates.

If A is and open subset in $\mathbb{R}$, then we may repeat the above argument with the circles $\mathcal{O}_\delta(z)$ 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 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

$$\lim_{x \to +\infty} \frac{a^x}{b^x} = 0, \ c \in \mathbb{R}, \quad 0 < \alpha < \beta, \quad 1 < a < b.$$

Hence we immediately have that

$$\lim_{x \to +\infty} \frac{3^{x+1} + x^5 - 4x}{3^x + 2^x + x^2} = \lim_{x \to +\infty} \frac{3 \cdot 3^x}{3^x} = 3,$$

$$\lim_{x \to +\infty} \frac{4^x - 8x^6 - 2^x - 167}{3^x - 45x - \sqrt{11}\pi^{x+12}} = \lim_{x \to +\infty} \frac{4^x}{-\sqrt{11}\pi^{12} \cdot \pi^x}$$
$$= -\infty.$$

If we realize that

$$\lim_{x \to +\infty} \operatorname{arctg} x = \frac{\pi}{2} \geq 1,$$

we will also obtain that

$$\lim_{x \to +\infty} \frac{\sqrt{x} - \sin^3 x + x \operatorname{arctg} x}{\sqrt{1 + 2x + x^2}} = \lim_{x \to +\infty} \frac{x \operatorname{arctg} x}{\sqrt{x^2}}$$
$$= \lim_{x \to +\infty} \operatorname{arctg} x = \frac{\pi}{2}.$$

$\square$

5.C.17. Determine the limits

$$\lim_{n \to \infty} \left(\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(n-1) \cdot n} \right);$$

$$\lim_{n \to \infty} \left(\frac{1}{\sqrt{n^2 + 1}} + \frac{1}{\sqrt{n^2 + 2}} + \cdots + \frac{1}{\sqrt{n^2 + n}} \right).$$

Solution. Since for every natural number $k \geq 2$ it holds that (what we do here is called partial fraction decomposition – we will present it in detail in the chapter concerning integration of rational functions)

$$\frac{1}{(k-1)k} = \frac{1}{k-1} - \frac{1}{k},$$

we get that

$$\lim_{n \to \infty} \left(\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(n-1) \cdot n} \right)$$
$$= \lim_{n \to \infty} \left(\frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{n-1} - \frac{1}{n} \right)$$
$$= \lim_{n \to \infty} \left(1 - \frac{1}{n} \right) = 1.$$

Let us remark that this limit is quite important: it determines the sum of one of the so-called telescoping series (with which Johann I Bernoulli (1667–1748) worked).

To determine the second limit, we invoke the squeeze theorem. The bounds

$$\frac{1}{\sqrt{n^2 + 1}} + \cdots + \frac{1}{\sqrt{n^2 + n}} \geq \frac{1}{\sqrt{n^2 + n}} + \cdots + \frac{1}{\sqrt{n^2 + n}}$$
$$= \frac{n}{\sqrt{n^2 + n}},$$

$$\frac{1}{\sqrt{n^2 + 1}} + \cdots + \frac{1}{\sqrt{n^2 + n}} \leq \frac{1}{\sqrt{n^2 + 1}} + \cdots + \frac{1}{\sqrt{n^2 + 1}}$$
$$= \frac{n}{\sqrt{n^2 + 1}}$$

real and imaginary projections of A, there is a subsequence $z_{n_k} = (x_{n_k}, y_{n_k})$ such that $x_{n_k} \to x$, $y_{n_k} \to y$ 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, $z_{n_k} \to z = (x, y)$, 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 circles replaced by closed intervals.

Finally, we have to prove the other implication. So assume $A \subset C$ 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

$$B_n = \{z = (\tfrac{k}{n}, \tfrac{m}{n}) \in A, \ k, m \in \mathbb{Z}\}$$

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 circles 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 must be at least one closed circle C in the system $\mathcal{A}_1$ which is not covered by a finite number the sets U_α. Call it C_1 and notice that diam $C_1 = 2\varepsilon_1$.

Next, consider the sets $C_1 \cap C$, with circles $C \in \mathcal{A}_2$ which cover the entire set C_1. Again, at least one of them cannot be covered by a finite number of U_α, we call it C_2. This way, we inductively construct a sequence of sets C_k satisfying $C_{k+1} \subset C_k$, diam $C_k \leq 2\varepsilon_k$, $\varepsilon_k \to 0$, and none of them can be covered by a finite number of the open sets U_α.

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_{α_0} containing z and containing also some δ-neighbourhood $\mathcal{O}_\delta(z)$. But now, if 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

298

for $n \in \mathbb{N}$ give that

$$\lim_{n\to\infty} \frac{n}{\sqrt{n^2+n}} \leq \lim_{n\to\infty} \left(\frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right)$$
$$\leq \lim_{n\to\infty} \frac{n}{\sqrt{n^2+1}}.$$

Since

$$\lim_{n\to\infty} \frac{n}{\sqrt{n^2+n}} = \lim_{n\to\infty} \frac{n}{\sqrt{n^2}} = 1,$$
$$\lim_{n\to\infty} \frac{n}{\sqrt{n^2+1}} = \lim_{n\to\infty} \frac{n}{\sqrt{n^2}} = 1,$$

we also have that

$$\lim_{n\to\infty} \left(\frac{1}{\sqrt{n^2+1}} + \frac{1}{\sqrt{n^2+2}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right) = 1.$$

$\square$

5.C.18. Calculate

(a)

$$\lim_{x\to 0} \frac{\sqrt{1+x}-\sqrt{1-x}}{x};$$

(b)

$$\lim_{x\to\pi/4} \frac{\cos x - \sin x}{\cos(2x)};$$

(c)

$$\lim_{x\to+\infty} \sqrt[3]{x^4} \left(\sqrt[3]{x^2+2x+3} - \sqrt[3]{x^2+2x+2} \right).$$

Solution. We will calculate the wanted limits using the method of multiplying both the numerator and the denominator by a suitable expression. The first fraction can be conveniently extended by

$$\sqrt{1+x} + \sqrt{1-x}$$

and making use of the well-known formula $(a-b)(a+b) = a^2 - b^2$. Thus we obtain

$$\lim_{x\to 0} \frac{\sqrt{1+x}-\sqrt{1-x}}{x} = \lim_{x\to 0} \frac{(1+x)-(1-x)}{x\left(\sqrt{1+x}+\sqrt{1-x}\right)}$$
$$= \lim_{x\to 0} \frac{2}{\sqrt{1+x}+\sqrt{1-x}} = \frac{2}{\sqrt{1}+\sqrt{1}} = 1.$$

Similarly we can calculate

$$\lim_{x\to\pi/4} \frac{\cos x - \sin x}{\cos(2x)} = \lim_{x\to\pi/4} \frac{(\cos x + \sin x)(\cos x - \sin x)}{(\cos x + \sin x)\cos(2x)}$$
$$= \lim_{x\to\pi/4} \frac{\cos^2 x - \sin^2 x}{(\cos x + \sin x)\cos(2x)}$$
$$= \lim_{x\to\pi/4} \frac{1}{\cos x + \sin x} = \frac{1}{\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}} = \frac{\sqrt{2}}{2}.$$

The reduction was made thanks to the identity

$$\cos(2x) = \cos^2 x - \sin^2 x, \quad x \in \mathbb{R}.$$

1-dimensional and the open circles are replaced by open intervals. $\square$

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,∞). Similarly, any interval $(-\infty, a)$ is a neighbourhood of $-\infty$. Further, we will extend the concept of a limit point so that ∞ is a limit point of a set $A \subset \mathbb{R}$ if and only if every neighbourhood of ∞ 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 $a \in \mathbb{R}$:

$$a + \infty = \infty$$
$$a - \infty = -\infty$$
$$a \cdot \infty = \infty, \text{ if } a > 0$$
$$a \cdot \infty = -\infty, \text{ if } a < 0$$
$$a \cdot (-\infty) = -\infty, \text{ if } a > 0$$
$$a \cdot (-\infty) = \infty, \text{ if } a < 0$$
$$\frac{a}{\pm\infty} = 0, \text{ for all } a \neq 0.$$

The following definition covers many cases of limit processes and needs to be thoroughly understood. Some particular cases are considered in great detail below.

Real and complex limits

Definition. Consider a subset $A \subset \mathbb{R}$ and a real-valued function $f : A \to \mathbb{R}$ or a complex-valued function $f : A \to \mathbb{C}$, 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

$$\lim_{x\to x_0} f(x) = a$$

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)!

Do we want to talk about deleted neighbourhood $\mathcal{O}(x) \setminus \{x\}$ of those points where we are interested in the function values?

As for the last limit, to make use of the formula

$$(a - b)\left(a^2 + ab + b^2\right) = a^3 - b^3,$$

we consider the expression

$$\sqrt[3]{\left(x^2 + 2x + 3\right)^2} + \sqrt[3]{x^2 + 2x + 3} \cdot \sqrt[3]{x^2 + 2x + 2}$$
$$+ \sqrt[3]{\left(x^2 + 2x + 2\right)^2},$$

which corresponds to $a^2 + ab + b^2$, so we choose

$$a = \sqrt[3]{x^2 + 2x + 3}, \qquad b = \sqrt[3]{x^2 + 2x + 2}.$$

By this extension, we transform the original limit to for some polynomials P, Q. Let us emphasize that this really holds for all $n \in \mathbb{N}$. For $n = 1$, we set $\binom{1}{2} = 0$ and the polynomials P, Q may be constant zeros. So we get for all real x:

$$(1 + 2nx)^n = 1 + 2n^2 x + 2n^3 (n - 1) x^2 + P(x) x^3,$$
$$(1 + nx)^{2n} = 1 + 2n^2 x + n^3 (2n - 1) x^2 + Q(x) x^3.$$

Mere substitution and simple rearrangements give us

$$\lim_{x \to 0} \frac{(1 + 2nx)^n - (1 + nx)^{2n}}{x^2} =$$
$$\lim_{x \to 0} \frac{\left(2n^3 (n - 1) - n^3 (2n - 1)\right) x^2 + (P(x) - Q(x)) x^3}{x^2}$$
$$= \lim_{x \to 0} \left(-n^3 + (P(x) - Q(x)) x\right) = -n^3 + 0 = -n^3.$$

$\square$

5.C.19. Calculate

$$\lim_{x \to \pi/4} (\tan x)^{\tan (2x)}.$$

Solution. Limits of the type $1^{\pm\infty}$ (like the examined one) can be calculated using the formula

$$\lim_{x \to x_0} f(x)^{g(x)} = e^{\lim_{x \to x_0} \left((f(x) - 1) g(x)\right)},$$

supposing the limit on the right-hand side exists and $f(x) \neq 1$ for all x of some deleted neighborhood of the point $x_0 \in \mathbb{R}$. Therefore, let us determine

$$\lim_{x \to \pi/4} (\tan x - 1) \tan (2x) = \lim_{x \to \pi/4} \left(\frac{\sin x}{\cos x} - 1\right) \frac{\sin (2x)}{\cos (2x)}$$
$$= \lim_{x \to \pi/4} \frac{\sin x - \cos x}{\cos x} \cdot \frac{2 \sin x \cos x}{\cos^2 x - \sin^2 x}$$
$$= \lim_{x \to \pi/4} \frac{-2 \sin x}{\cos x + \sin x} = \frac{-2 \frac{\sqrt{2}}{2}}{\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}} = -1.$$

Hence we have that

$$\lim_{x \to \pi/4} (\tan x)^{\tan (2x)} = \frac{1}{e}.$$

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 ∞, and we mostly write the values (terms) of the sequence as $f(n) = a_n$ and the limit in the form

$$\lim_{n \to \infty} a_n = a.$$

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 $\mathrm{Re}\, a$ and the imaginary parts converge to $\mathrm{Im}\, a$.

(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

$$\lim_{x \to x_0} f(x) = a.$$

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 $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ 1 & \text{if } x = 0. \end{cases}$$

Apparently, the limit at zero is well-defined, and in accordance with our expectations, $\lim_{x \to 0} f(x) = 0$ even though the value $f(0) = 1$ does not belong into small neighbourhoods of the limit point 0.

An equivalent definition using ε-neighbourhoods of the limits a and δ-neighbourhoods of the limit points x_0 is the following: $\lim_{x \to x_0} f(x) = a$ *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 $\lim_{x \to x_0^+} f(x)$ and $\lim_{x \to x_0^-} f(x)$, respectively. As an example, we can consider the one-sided limits at $x_0 = 0$ for Heaviside's function

Let us remark that the used formula holds more generally for "the type 1^{whatever}", that is with no further conditions on the limit $\lim_{x \to x_0} g(x)$ which even need not exist. □

5.C.20. Show that
$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

Solution. Let us consider the unit circle (especially its quarter lying in the first quadrant) and its point $[\cos x, \sin x]$, $x \in (0, \pi/2)$. The length of the arc between the points $[\cos x, \sin x]$ and $[1, 0]$ is equal to x. So we apparently have
$$\sin x < x, \quad x \in \left(0, \frac{\pi}{2}\right).$$

The value $\tan x$ is then the distance between the points $[1, \sin x / \cos x]$ and $[1, 0]$. We can see that (feel free to draw a picture)
$$x < \tan x, \quad x \in \left(0, \frac{\pi}{2}\right).$$

This inequality also follows from the fact that the area of the triangle with vertices $[0, 0]$, $[1, 0]$, $[1, \tan x]$ is greater than the area of the considered circular sector. Altogether, we have obtained that
$$\sin x < x < \frac{\sin x}{\cos x}, \quad x \in \left(0, \frac{\pi}{2}\right),$$
that is
$$1 < \frac{x}{\sin x} < \frac{1}{\cos x}, \quad x \in \left(0, \frac{\pi}{2}\right),$$
$$1 > \frac{\sin x}{x} > \cos x, \quad x \in \left(0, \frac{\pi}{2}\right).$$

Invoking the squeeze theorem, we get the inequalities
$$1 = \lim_{x \to 0+} 1 \geq \lim_{x \to 0+} \frac{\sin x}{x} \geq \lim_{x \to 0+} \cos x = \cos 0 = 1.$$

Thus we have proved that
$$\lim_{x \to 0+} \frac{\sin x}{x} = 1.$$

The function $y = (\sin x)/x$ defined for $x \neq 0$ is even, whence it follows that
$$\lim_{x \to 0-} \frac{\sin x}{x} = \lim_{x \to 0+} \frac{\sin x}{x} = 1.$$

Since both one-sided limits exist and have the same value, the examined limit exists as well and satisfies
$$\lim_{x \to 0} \frac{\sin x}{x} = \lim_{x \to 0\pm} \frac{\sin x}{x} = 1.$$

Let us remark that at first sight, one could say the limit can be calculated using l'Hospital's rule. However, then one would have to know the sine's derivative at zero which, actually, is the limit in question. Thus we may not invoke l'Hospital's rule in this case. □

h from the beginning of this part. Apparently,
$$\lim_{x \to 0+} h(x) = 1, \quad \lim_{x \to 0-} h(x) = 0.$$
However, the limit $\lim_{x \to 0} f(x)$ 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 $f : A \to \mathbb{C}$ 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
$$\lim_{x \to x_0} f(x) = \lim_{x \to x_0} (\operatorname{Re} f(x)) + i \lim_{x \to x_0} (\operatorname{Im} f(x)).$$

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 δ–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 $x \in \mathbb{R}$,
$$\lim_{x \to x_0} f(x) = f(x_0).$$
Really, if $f(x) = a_n x^n + \cdots + a_0$, then the identity $(x_0 + \delta)^k = x_0^k + k\delta x_0^{k-1} + \cdots + \delta^k$, substituted for $k = 0, \ldots, n$, gives that choosing a sufficiently small δ 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.

(4) The following function is even trickier than the previous one. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as follows:[8]
$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{q} & \text{if } x = \frac{p}{q} \in \mathbb{Q}, \ p, q > 0 \text{ are co-prime,} \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

---
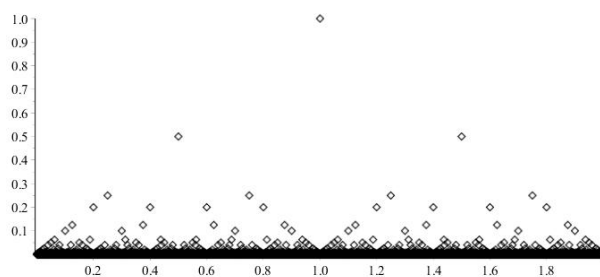
[8]This function is called the *Thomae function* after the German mathematician J. Thomae, 1840–1921. You may find it under many other names too: e.g. *Riemann function, pop-corn function, raindrop function* etc. It illustrates how badly dense the "discontinuity" points of a function can be even though it has limits everywhere.

5.C.21. Determine the limits

$$\lim_{n\to\infty}\left(\frac{n}{n+1}\right)^n,\ \lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^n,\ \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n^2},$$

$$\lim_{x\to0}\frac{\sin^2 x}{x},\qquad \lim_{x\to0}\frac{x}{\sin^2 x},\qquad \lim_{x\to0}\frac{\arcsin x}{x},$$

$$\lim_{x\to0}\frac{3\tan^2 x}{5\,x^2},\qquad \lim_{x\to0}\frac{\sin(3x)}{\sin(5x)},\qquad \lim_{x\to0}\frac{\tan(3x)}{\sin(5x)},$$

$$\lim_{x\to0}\frac{e^{5x}-e^{2x}}{x},\qquad \lim_{x\to0}\frac{e^{5x}-e^{-x}}{\sin(2x)}.$$

Solution. When calculating these limits, we will use our knowledge of the following limits ($a\in\mathbb{R}$):

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n=e^a,\ \lim_{x\to0}\frac{\sin x}{x}=1,\ \lim_{x\to0}\frac{e^x-1}{x}=1.$$

Thus we know that

$$e^{-1}=\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n=\lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n.$$

The substitution $m=n-1$ gives us

$$\lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n=\lim_{m\to\infty}\left(\frac{m}{m+1}\right)^{m+1}$$

$$=\lim_{m\to\infty}\left(\frac{m}{m+1}\right)^m\cdot\lim_{m\to\infty}\frac{m}{m+1}.$$

Altogether, we have

$$e^{-1}=\lim_{m\to\infty}\left(\frac{m}{m+1}\right)^m\cdot\lim_{m\to\infty}\frac{m}{m+1}.$$

Clearly, the second limit is equal to 1. Changing the variables (replacing n with m), we can write the result

$$e^{-1}=\lim_{n\to\infty}\left(\frac{n}{n+1}\right)^n.$$

Further, it holds that

$$\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^n=\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^{\frac{n^2}{n}}$$

$$=\lim_{n\to\infty}\left(\left(1+\frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}}=e^0=1$$

and

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n^2}=\lim_{n\to\infty}\left(\left(1-\frac{1}{n}\right)^n\right)^n=0.$$

Let us point out that the first result follows from the limits

$$\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^{n^2}=\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m=e,\ \lim_{n\to\infty}\frac{1}{n}=0$$

and the second one from

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n=e^{-1},\qquad \lim_{n\to\infty}n=+\infty,$$

where $e^{-\infty}=0\ (=\lim_{x\to-\infty}e^x=0)$.

Choose any point x_0, no matter whether rational or irrational. Our goal is show that $\lim_{x\to x_0}f(x)=0$. Thus fix any $\varepsilon>0$ and look at the possible values of $f(x)$ close to x_0. Notice that $f(x)\ge\varepsilon$, i.e. $\frac{1}{q}\ge\varepsilon$ 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 $f(x)\ge\varepsilon$. Label them $x_1,\dots,x_n$. Finally, choose δ 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 deleted neighbourhood of a limit point $x_0\in\mathbb{R}$ of the domain where*
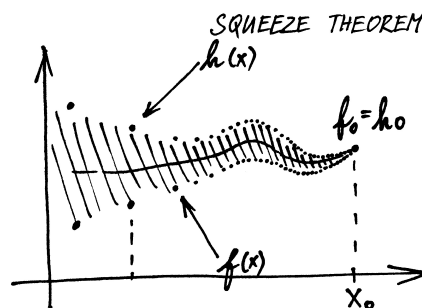
$$f(x)\le g(x)\le h(x).$$

Suppose there are limits

$$\lim_{x\to x_0}f(x)=f_0,\qquad \lim_{x\to x_0}h(x)=h_0$$

and $f_0=h_0$. Then the limit

$$\lim_{x\to x_0}g(x)=g_0$$

exists, and it satisfies $g_0=f_0=h_0$.

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 $x\ne x_0$. From the condition $f(x)\le g(x)\le h(x)$, it follows that $g(x)\in(f_0-\varepsilon,f_0+\varepsilon)$, and so $\lim_{x\to x_0}g(x)=f_0$.

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 ε. 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

We can easily get that

$$\lim_{x \to 0} \frac{\sin^2 x}{x} = \lim_{x \to 0} \sin x \cdot \lim_{x \to 0} \frac{\sin x}{x} = 0 \cdot 1 = 0.$$

Apparently,

$$\lim_{x \to 0} \frac{x}{\sin x} = 1^{-1} = 1$$

and the limit

$$\lim_{x \to 0} \frac{1}{\sin x}$$

does not exist (we write $1/\pm 0$). If we used the rule for the limit of a product to determine the limit

$$\lim_{x \to 0} \frac{x}{\sin^2 x},$$

we would obtain $1 \cdot 1/\pm 0 = 1/\pm 0$. This means that the limit does not exist (this, again, is a determinate form). For the calculation of

$$\lim_{x \to 0} \frac{\arcsin x}{x},$$

we will make use of the identity $x = \sin(\arcsin x)$ which holds for any $x \in (-1, 1)$, that is in some neighborhood of the point 0. Substituting $y = \arcsin x$, we get

$$\lim_{x \to 0} \frac{\arcsin x}{x} = \lim_{x \to 0} \frac{\arcsin x}{\sin(\arcsin x)} = \lim_{y \to 0} \frac{y}{\sin y} = 1.$$

Let us remark that $y \to 0$ 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).

We can immediately see that

$$\begin{aligned}
\lim_{x \to 0} \frac{3 \tan^2 x}{5 \, x^2} &= \lim_{x \to 0} \left(\frac{3}{5} \cdot \frac{\sin x}{x} \cdot \frac{\sin x}{x} \cdot \frac{1}{\cos^2 x} \right) \\
&= \frac{3}{5} \cdot \lim_{x \to 0} \frac{\sin x}{x} \cdot \lim_{x \to 0} \frac{\sin x}{x} \cdot \lim_{x \to 0} \frac{1}{\cos^2 x} \\
&= \frac{3}{5} \cdot 1 \cdot 1 \cdot 1 = \frac{3}{5}.
\end{aligned}$$

By appropriate extension and substitution, we get

$$\begin{aligned}
\lim_{x \to 0} \frac{\sin(3x)}{\sin(5x)} &= \lim_{x \to 0} \left(\frac{\sin(3x)}{3x} \cdot \frac{5x}{\sin(5x)} \cdot \frac{3}{5} \right) \\
&= \lim_{x \to 0} \frac{\sin(3x)}{3x} \cdot \lim_{x \to 0} \frac{5x}{\sin(5x)} \cdot \frac{3}{5} \\
&= \lim_{y \to 0} \frac{\sin y}{y} \cdot \lim_{z \to 0} \frac{z}{\sin z} \cdot \frac{3}{5} = 1 \cdot 1 \cdot \frac{3}{5} = \frac{3}{5}.
\end{aligned}$$

Thanks to the previous result, it can easily be calculated that

$$\begin{aligned}
\lim_{x \to 0} \frac{\tan(3x)}{\sin(5x)} &= \lim_{x \to 0} \left(\frac{\sin(3x)}{\sin(5x)} \cdot \frac{1}{\cos(3x)} \right) \\
&= \lim_{x \to 0} \frac{\sin(3x)}{\sin(5x)} \cdot \lim_{x \to 0} \frac{1}{\cos(3x)} = \frac{3}{5} \cdot 1 = \frac{3}{5}.
\end{aligned}$$

interval (M, ∞). 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*

$$\lim_{x \to x_0} f(x) = a \in \mathbb{K}, \quad \lim_{x \to x_0} g(x) = b \in \mathbb{K}$$

exist. Then:
(1) the limit a is unique,
(2) the limit of the sum $f + g$ exists and satisfies

$$\lim_{x \to x_0} (f(x) + g(x)) = a + b,$$

(3) the limit of the product $f \cdot g$ exists and satisfies

$$\lim_{x \to x_0} (f(x) \cdot g(x)) = a \cdot b.$$

In particular, if $f(x) = a$ is a constant function then $\lim_{x \to x_0} a \cdot g(x) = a \cdot b$,
(4) if $b \neq 0$, the limit of the quotient f/g exists and satisfies

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{a}{b}.$$

PROOF. (1) Suppose a and a' are two values of the limit $\lim_{x \to x_0} f(x)$. 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 $x \neq x_0$, both $f(x)$ and $g(x)$ will lie in ε–neighbourhoods of the points a and b. Hence their sum will lie in the 2ε–neighbourhood of $a + b$. The proposition is proved.

(3) Similarly to the above paragraph: choose $\mathcal{O}_{\varepsilon^2}(ab)$. For sufficiently small neighbourhoods of x_0, the values of both f and g will lie in ε–neighbourhoods of the values a and b. Therefore, their product will lie in the required ε^2–neighbourhood.

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 $\lim_{x \to x_0} \frac{1}{g(x)} = \frac{1}{b}$ 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. $\square$

Similarly, we can determine

$$\lim_{x \to 0} \frac{e^{5x} - e^{2x}}{x} = \lim_{x \to 0} \left(e^{2x} \frac{e^{(5-2)x} - 1}{(5-2)x} (5 - 2) \right)$$

$$= \lim_{x \to 0} e^{2x} \cdot \lim_{x \to 0} \frac{e^{3x} - 1}{3x} \cdot 3$$

$$= e^0 \cdot \lim_{y \to 0} \frac{e^y - 1}{y} \cdot 3 = 1 \cdot 1 \cdot 3 = 3$$

and also

$$\lim_{x \to 0} \frac{e^{5x} - e^{-x}}{\sin(2x)} = \lim_{x \to 0} \left(\frac{e^{5x} - 1}{\sin(2x)} - \frac{e^{-x} - 1}{\sin(2x)} \right)$$

$$= \lim_{x \to 0} \left(\frac{e^{5x} - 1}{5x} \cdot \frac{2x}{\sin(2x)} \cdot \frac{5}{2} \right.$$

$$\left. - \frac{e^{-x} - 1}{-x} \cdot \frac{2x}{\sin(2x)} \cdot \left(-\frac{1}{2} \right) \right)$$

$$= \lim_{x \to 0} \frac{e^{5x} - 1}{5x} \cdot \lim_{x \to 0} \frac{2x}{\sin(2x)} \cdot \frac{5}{2}$$

$$- \lim_{x \to 0} \frac{e^{-x} - 1}{-x} \cdot \lim_{x \to 0} \frac{2x}{\sin(2x)} \cdot \left(-\frac{1}{2} \right)$$

$$= \frac{5}{2} \lim_{u \to 0} \frac{e^u - 1}{u} \cdot \lim_{z \to 0} \frac{z}{\sin z} + \frac{1}{2} \lim_{v \to 0} \frac{e^v - 1}{v} \cdot \lim_{z \to 0} \frac{z}{\sin z}$$

$$= \frac{5}{2} + \frac{1}{2} = 3.$$

□

5.C.22. Calculate the limits

$$\lim_{x \to 0} \frac{1 - \cos(2x)}{x \sin x}; \qquad \lim_{x \to 0} \frac{1 - \cos x}{x^2}.$$

Solution. We will utilize the fact that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

Then we get

$$\lim_{x \to 0} \frac{1 - \cos(2x)}{x \sin x} = \lim_{x \to 0} \frac{1 - (\cos^2 x - \sin^2 x)}{x \sin x}$$

$$= \lim_{x \to 0} \frac{(1 - \cos^2 x) + \sin^2 x}{x \sin x}$$

$$= \lim_{x \to 0} \frac{2 \sin^2 x}{x \sin x} = \lim_{x \to 0} 2 \frac{\sin x}{x} = 2;$$

and

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \lim_{x \to 0} \left(\frac{1 - \cos x}{x^2} \cdot \frac{1 + \cos x}{1 + \cos x} \right)$$

$$= \lim_{x \to 0} \frac{1 - \cos^2 x}{x^2 (1 + \cos x)} = \lim_{x \to 0} \frac{\sin^2 x}{x^2 (1 + \cos x)}$$

$$= \left(\lim_{x \to 0} \frac{\sin x}{x} \right)^2 \cdot \lim_{x \to 0} \frac{1}{1 + \cos x} = \frac{1}{2}.$$

Let us remark that we could also use the identity

$$1 - \cos(2x) = 2 \sin^2 x, \qquad x \in \mathbb{R}.$$

□

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 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
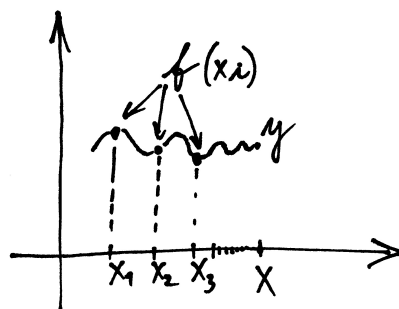
PROOF. Suppose first that the limit of f at x_0 is y. Then for any neighbourhood U of the point y, there is a neighbourhood V of x_0 such that for all $x \in V \cap A$, $x \neq x_0$, $f(x) \in U$. For every sequence $x_n \to x_0$ of points different from x_0, the terms x_n lie in V for all n greater 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 $x \to x_0$. 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 $\lim_{m \to \infty} x_m = x_0$, for which the values $f(x_n)$ do not converge to y. The proof is finished. □

D. Continuity of functions

5.D.1. Let us examine existence of limits and continuity of the function $(x-1)^{-\operatorname{sgn} x}$ at the points 0 and 1.

Solution. First, let us calculate the one-sided limits at the point 0:

$$\lim_{x \to 0^-} (x-1)^{-\operatorname{sgn} x} = \lim_{x \to 0^-} (x-1) = -1,$$
$$\lim_{x \to 0^+} (x-1)^{-\operatorname{sgn} x} = \lim_{x \to 0^+} \frac{1}{x-1} = -1,$$

whence $\lim_{x \to 0} (x-1)^{-\operatorname{sgn} x} = -1$. However, the function value at 0 equals 1, so the examined function is not continuous at the point 0. Further, we have that

$$\lim_{x \to 1^-} (x-1)^{-\operatorname{sgn} x} = \lim_{x \to 1^-} \frac{1}{x-1} = -\infty,$$
$$\lim_{x \to 1^+} (x-1)^{-\operatorname{sgn} x} = \lim_{x \to 1^+} \frac{1}{x-1} = \infty.$$

Both one-sided limits at the point 1 exist, yet they differ, which implies that the (two-sided) limit of this function at 1 does not exist, and the function is not continuous here, either. □

5.D.2. Without invoking the squeeze theorem, prove that the function

$$R(x) = \begin{cases} x, & x \in \left\{ \frac{1}{n}; \, n \in \mathbb{N} \right\}; \\ 0, & x \in \mathbb{R} \smallsetminus \left\{ \frac{1}{n}; \, n \in \mathbb{N} \right\} \end{cases}$$

is continuous at the point 0.

Solution. The function R is continuous at the point 0 if and only if

$$\lim_{x \to 0} R(x) = R(0) = 0.$$

We will show that, by the definition of a limit, the examined limit equals 0. Let $\delta > 0$ be arbitrary. For any $x \in (-\delta, \delta)$ we have that $R(x) = 0$, or $R(x) = x$, hence (in both cases) we get $R(x) \in (-\delta, \delta)$. In other words, having chosen any δ-neighborhood $(-\delta, \delta)$ of the point a, we can take the δ-neighborhood $(-\delta, \delta)$ of the point x_0 as then for any $x \in (-\delta, \delta)$ (the considered neighborhood of x_0) it holds that $R(x) \in (-\delta, \delta)$ (here, the interval $(-\delta, \delta)$ is the neighborhood of a). This matches the definition of a limit (we did not even have to require $x \neq x_0$).

The considered function R is called the Riemann function (hence the name R). In literature, it can be found in many modifications. For instance, the function

$$f(x) = \begin{cases} 1, & x \in \mathbb{Z}; \\ \frac{1}{q}, & x = \frac{p}{q} \in \mathbb{Q}; \, p, q \text{ relatively prime}, q > 1; \\ 0, & x \notin \mathbb{Q} \end{cases}$$

is also "often" called the Riemann function. □

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

$$\lim_{x \to x_0} f(x) = f(x_0).$$

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 ε-neighbourhood $\mathcal{O}_\varepsilon(f(x_0))$ is shown), there is a neighbourhood of x_0 (the δ-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.
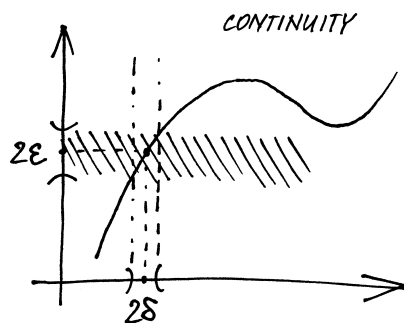
Notice that for the boundary points of the interval A, the definition says that value of f equals the value of the one-sided limit there. The function is said to be *right-continuous* or *left-continuous* at such a point. Every polynomial is a continuous function on the whole $\mathbb{R}$, see 5.2.11(2). 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

5.D.3. By defining the values at the points -1 and 1, extend the function

$$f(x) = (x^2 - 1)\sin\frac{2x-1}{x^2-1}, \quad x \neq \pm 1 \, (x \in \mathbb{R})$$

so that the resulting function is continuous on the whole $\mathbb{R}$.

Solution. The original function is continuous at every point of its domain. Thus the extended function will be continuous if and only if we set

$$
\begin{aligned}
f(-1) &:= \lim_{x\to -1}\left((x^2-1)\sin\frac{2x-1}{x^2-1}\right), \\
f(1) &:= \lim_{x\to 1}\left((x^2-1)\sin\frac{2x-1}{x^2-1}\right).
\end{aligned}
$$

If either of these limits did not exist (or were infinite), the function could not be extended to a continuous one. Clearly we have that

$$\left|\sin\frac{2x-1}{x^2-1}\right| \leq 1, \quad x \neq \pm 1 \, (x \in \mathbb{R}),$$

whence it follows that

$$-\left|x^2-1\right| \leq f(x) \leq \left|x^2-1\right|, \quad x \neq \pm 1 \, (x \in \mathbb{R}).$$

Since

$$\lim_{x\to\pm 1}\left|x^2-1\right| = 0,$$

by the squeeze theorem, we get the result $f(\pm 1) := 0$. □

5.D.4. 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

$$f(0) = -4, \quad \lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} e^{2x} = +\infty.$$

From the fact that f is continuous on the whole domain it thus follows that it takes on all values $y \in [-4, +\infty)$. Especially, its graph necessarily intersects the positive semiaxis x, ie. the equation $f(x) = 0$ has a solution. □

5.D.5. At which points $x \in \mathbb{R}$ is the function

$$y = \cos\left(\operatorname{arctg}\left(\left|12x^{21} + 11\right| \cdot \frac{e^{\cos(x+2)-x^3}}{-11-x^{12}}\right)\right) + \sin(\sin(\sin x))$$

(considering maximum domain) continuous? ○

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 ε–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 δ-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. □

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.

TOPOLOGICAL CHARACTERIZATION OF CONTINUITY

Theorem. *Let $f : A \subset \mathbb{R} \to \mathbb{R}$ be a continuous 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,

(2) the inverse image $f^{-1}(W)$ of every closed set $W \subset \mathbb{R}$ is a closed set,

(3) the image $f(K)$ of every compact set $K \subset A$ is a compact set,

(4) 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, which finishes the proof.

Conversely, if $f^{-1}(U)$ is open for each open U, then taking any ε-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 W.

(3) Choose any open cover of $f(K)$. The inverse images of the particular intervals are unions of open intervals and thus create a cover of the set K. Select a finite subcover from it. Then finitely many of the corresponding images cover the original set $f(K)$.

5.D.6. Determine whether the function

$$f(x) = \begin{cases} x, & x < 0; \\ 0, & 0 \le x < 1; \\ x, & x = 1; \\ 0, & 1 < x < 2; \\ x, & 2 \le x \le 3; \\ \frac{1}{x-3}, & x > 3 \end{cases}$$

is continuous; left-continuous; right-continuous at the points $-\pi, 0, 1, 2, 3, \pi$. ○

5.D.7. Extend the function

$$f(x) = \operatorname{arctg}\left(1 + \frac{5}{x^2}\right) \cdot \sin^2 x^5, \qquad x \in \mathbb{R} \smallsetminus \{0\}$$

at $x = 0$ so that it is continuous at this point. ○

5.D.8. Find all $p \in \mathbb{R}$ for which the function

$$f(x) = \frac{\sin(6x)}{3x}, \quad x \in \mathbb{R} \smallsetminus \{0\}; \qquad f(0) = p$$

is continuous at the origin. ○

5.D.9. Choose a real number a so that the function

$$h(x) = \frac{x^4 - 1}{x - 1}, \quad x > 1; \qquad h(x) = a, \quad x \le 1$$

is continuous on $\mathbb{R}$. ○

5.D.10. Calculate

$$\lim_{x \to 0+} \frac{\sin^8 x}{x^3}; \qquad \lim_{x \to -\infty} \frac{\sin^8 x}{x^3}.$$

○

5.D.11. Find all possible values of the parameter $a \in \mathbb{R}$ so that the inequality

$$(a-2)x^2 - (a-2)x + 1 > 0$$

holds for all real numbers x.

Solution. We can notice that for $a = 2$, the inequality holds trivially (there is constant 1 on the left side). For $a \ne 2$, the left side is a quadratic function $f(x)$ in the variable x, and further $f(0) = 1$. Thanks to the function $f(x)$ being continuous, the inequality $f(x) > 0$ will hold for all real x if and only if there is no solution to the equation $f(x) = 0$ in $\mathbb{R}$ (the whole of the graph of the function f will then be "above" the x-axis). This will occur if and only if the discriminant of the quadratic equation $(a-2)x^2 - (a-2)x + 1 = 0$ (in x) will be negative. Thus we get the following necessary and sufficient condition:

$$D = (a-2)^2 - 4(a-2) = (a-2)(a-6) < 0.$$

This is true for $a \in (2, 6)$. Altogether, the inequality holds for all real x iff $a \in [2, 6)$. □

(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. □

5.2.19. There are two very useful consequences of the previous theorem.

MAXIMA AND MINIMA OF CONTINUOUS FUNCTIONS[9]

Corollary. *Let $f : \mathbb{R} \to \mathbb{R}$ 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. □

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.

---

[9]This result is usually called the *Weierstrass theorem*, but it is also known (especially in Czech literature) as the Bolzano's theorem. Bernard Bolzano worked in Prague at the beginning of the 19th century and apparently used such the result as a technical lemma when proving his Bolzano-Weierstrass theorem mentioned earlier.

5.D.12. In $\mathbb{R}$, solve the equation

$$2^x + 3^x + 4^x + 5^x + 6^x = 5.$$

Solution. The function on the left side is a sum of five increasing functions on $\mathbb{R}$, so it must be increasing as well. For $x = 0$, its value is 5, which is thus the only solution of the equation. $\square$

5.D.13. In $\mathbb{R}$, solve the equation
$2^x + 3^x + 6^x = 1$. $\bigcirc$

5.D.14. Determine whether the polynomial

$$P(x) = x^{37} + 5x^{21} - 4x^9 + 5x^4 - 2x - 3$$

has a real root in the interval $(-1, 1)$. $\bigcirc$

E. Derivatives

First of all, let us show that the derivatives enlisted in the table of paragraph 5.3.1 are correct. We will derive them right from the definition of a derivative.

5.E.1. From the definition, (see 5.3.1) find the derivatives of the functions x^n (x is the variable, n is a constant positive integer), $\sqrt{x}$, $\sin x$.

Solution. First, let us remark that by substituting h for $x - x_0$ in the definition of a derivative, we get

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

In the following calculations, we will work with the latter expression of the limit.

$$(x^n)' = \lim_{h \to 0} \frac{(x+h)^n - x^n}{h}$$
$$= \lim_{h \to 0} \frac{\binom{n}{1} x^{n-1} h + \binom{n}{2} x^{n-2} h^2 + \cdots + h^n}{h} =$$
$$nx^{n-1} + \lim_{h \to 0} \left(\binom{n}{2} x^{n-2} h + \binom{n}{3} x^{n-3} h^2 + \cdots + h^{n-1} \right)$$
$$= nx^{n-1},$$

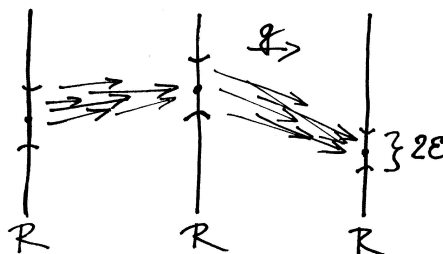**Theorem.** *Let f, g : $\mathbb{R} \to \mathbb{R}$ be two functions and $\lim_{x \to a} f(x) = b$.*

(1) If the function g is continuous at the point b, then

$$\lim_{x \to a} g\left(f(x)\right) = g\left(\lim_{x \to a} f(x)\right) = g(b).$$

(2) If the limit $\lim_{y \to b} g(y)$ exists and $f(x) \neq b$ holds for all $x \neq a$ from some neighbourhood of the point a, then

$$\lim_{x \to a} g\left(f(x)\right) = \lim_{y \to b} g(y).$$

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. $\square$

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 $h : \mathbb{R} \setminus \{x \in \mathbb{R}, \ g(x) = 0\} \to \mathbb{C}$,

$$h(x) = \frac{f(x)}{g(x)}$$

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 they are undefined, they can have

- a finite limit,

$$(\sqrt{x})' = \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h}$$
$$= \lim_{h \to 0} \frac{(\sqrt{x+h} - \sqrt{x})(\sqrt{x+h} + \sqrt{x})}{h(\sqrt{x+h} + \sqrt{x})}$$
$$= \lim_{h \to 0} \frac{h}{h(\sqrt{x+h} + \sqrt{x})} = \lim_{h \to 0} \frac{1}{\sqrt{x+h} + \sqrt{x}}$$
$$= \frac{1}{2\sqrt{x}},$$

$$(\sin x)' = \lim_{h \to 0} \frac{\sin(x+h) - \sin x}{h}$$
$$= \lim_{h \to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h}$$
$$= \lim_{h \to 0} \frac{\cos x \sin h}{h} + \lim_{h \to 0} \frac{\sin x(\cos h - 1)}{h}$$
$$= \cos x \cdot \lim_{h \to 0} \frac{\sin h}{h} - \lim_{h \to 0} \frac{2(\sin \frac{h}{2})^2}{h}$$
$$= \cos x \cdot 1 + \lim_{t \to 0} \sin t \frac{\sin t}{t}$$
$$= \cos x.$$

$\square$

5.E.2. Differentiate:

i) $x \sin x$,

ii) $\frac{\sin x}{x}$, $x \neq 0$

iii) $\ln(x + \sqrt{x^2 - a^2})$, $a \neq 0$, $|x| \geq |a|$,

iv) $\arctan\left(\frac{x}{\sqrt{1-x^2}}\right)$, $|x| < 1$,

v) x^x, $x > 0$.

Solution. (i) By the formula for the derivative of a product (the Leibniz rule, see 5.3.4) we get

$$(x \sin x)' = x' \cdot \sin x + x \cdot (\sin x)' = \sin x + x \cos x.$$

(ii) By the formula for the derivative of a quotient (5.3.5), we have that

$$\frac{\sin x}{x} = \frac{(\sin x)' \cdot x - \sin x \cdot x'}{x^2} = \frac{x \cos x - \sin x}{x^2}.$$

(iii) This time, we will use the formula for the derivative of function composition (the chain rule, see 5.3.4). Setting $h(x) = \ln(x)$, $f(x) = x + \sqrt{x^2 - a^2}$, we obtain

$$\ln(x + \sqrt{x^2 - a^2})' = h(f(x))' = h'(f(x)) \cdot f'(x)$$
$$= \frac{(x + \sqrt{x^2 - a^2})'}{x + \sqrt{x^2 - a^2}} = \frac{1 + \frac{x}{x^2 - a^2}}{x + \sqrt{x^2 - a^2}},$$

where we used the chain rule once again when differentiating $\sqrt{x^2 - a^2}$.

- 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 function's domain by this point can be extended by defining it to take the value of the limit there. Then the function is continuous at the same point.

The possibilities are illustrated in the diagram showing the values of the function

$$h(x) = \frac{(x - 0.05a)(x - 2 - 0.2a)(x - 5)}{x(x - 2)(x - 4)}$$

for $a = 0$ on the left (thus it is the rational function $(x - 5)/(x - 4)$) and for $a = 5/3$ on the right.
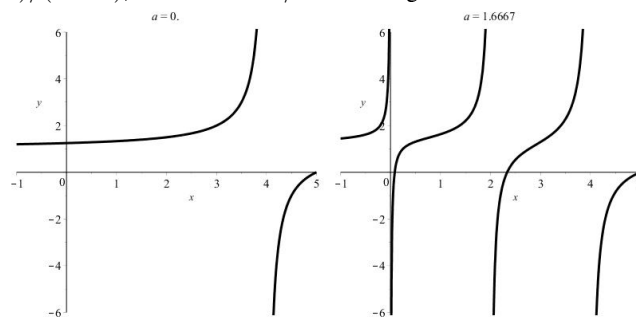
5.2.22. Power functions and the exponential. We have met the simple power functions $x \mapsto x^n$ with natural exponents $n = 0, 1, 2, \ldots$ when building the polynomials. The meaning of the function $x \mapsto x^{-1}$, defined for all $x \neq 0$, is also clear. Now, we extend this definition to a general *power function* x^a with an arbitrary $a \in \mathbb{R}$.

We use the well known properties of powers and roots. For a negative integer $-a$, we define

$$x^{-a} = (x^a)^{-1} = (x^{-1})^a.$$

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. $b = x^{\frac{1}{n}}$. 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, 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, $b^n = x$. 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

$$x^a = \sup\{x^y, y \in \mathbb{Q}, \ y \leq a\}.$$

(iv) Again, we are looking for the derivative of a composed function:

$$\left[\arctan\left(\frac{x}{\sqrt{1-x^2}}\right)\right]' = \frac{1}{1+\frac{x^2}{1-x^2}} \cdot \left(\frac{x}{\sqrt{1-x^2}}\right)'$$

$$= \frac{1}{1+\frac{x^2}{1-x^2}} \cdot \frac{\sqrt{1-x^2}+\frac{x^2}{\sqrt{1-x^2}}}{1-x^2}$$

$$= \sqrt{1-x^2} + \frac{x^2}{\sqrt{1-x^2}} = \frac{1}{\sqrt{1-x^2}}.$$

(v) First, we transform the function to a function with constant base (preferably the base e) which we are able to differentiate.

$$(x^x)' = \left((e^{\ln x})^x\right)' = (e^{x\ln x})'$$

$$= (x\ln x)' \cdot e^{x\ln x} = (1+\ln x)\cdot x^x$$

$\square$

5.E.3. Find the derivative of the function $y = x^{\sin x}$, $x > 0$.

Solution. We have

$$\left(x^{\sin x}\right)' = \left(e^{\sin x \, \ln x}\right)' = e^{\sin x \, \ln x}\left(\cos x \, \ln x + \frac{\sin x}{x}\right)$$

$$= x^{\sin x}\left(\cos x \, \ln x + \frac{\sin x}{x}\right).$$

$\square$

5.E.4. For positive x, determine the derivative of the function $f(x) = x^{\ln x}$ $\bigcirc$

5.E.5. For $x \in (0, \pi/2)$, calculate the derivative of the function $y = (\sin x)^{\cos x}$. $\bigcirc$

We advise the reader to make up some functions and find their derivatives. The results can be verified in a great deal of mathematical programs. In the following exercise, we will look at the geometrical meaning of the derivative at a given point, namely that it determines the slope of the tangent line to the function graph at the given point (see 5.3.2).

5.E.6. Using differential, approximate arccotg 1.02.

Solution. The differential of the function f having continuous derivative at the point x_0 is equal to

$$f'(x_0)\, dx = f'(x_0)\,(x-x_0).$$

The equation of the tangent to f's graph at the point $[x_0, f(x_0)]$ is then

$$y - f(x_0) = f'(x_0)\,(x-x_0).$$

Hence we can see that the differential is the growth on the tangent line. However, the values on the tangent approximate those of f, supposing the difference $x-x_0$ is "small". Thus

As before, $x^{-a} = \frac{1}{x^a}$.

For $0 < x < 1$, proceed analogously with care for the inequality signs, or set $x^a = (\frac{1}{x})^{-a}$. 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 $a \in \mathbb{R}$. There is another view of the construction: For every fixed real number $c > 0$, there is a well-defined function $y \mapsto c^y$ on the whole real line. This function is called the *exponential function* with base c. The definition ensures that the resulting function c^y is continuous in y for fixed c and is continuous in c for fixed y.

The properties used when defining the power function and the exponential function $f(y) = c^y$, with $f(0) = 1$, can be summarized in a single inequality for any real x and real y:

$$(1) \qquad f(x+y) = f(x)\cdot f(y)$$

together with the condition of continuity.

Indeed, for $y = 0$ $f(0) = 1$, and hence $1 = f(0) = f(x-x) = f(x)\cdot(f(x))^{-1}$. Eventually, for a natural number n, $f(nx) = (f(x))^n$. Thus a^x is determined for all $a > 0$ and $x \in \mathbb{Q}$. The continuity condition determines the function's values uniquely at the remaining points, if such a function exists. So we should check, that the above constructed function has the property 1. This should be clear from the continuity of the operations of sum and product in both arguments – the details are left to the reader.

Thus, the exponential function satisfies both the well-known formulas

$$(2) \qquad a^x \cdot a^y = a^{x+y}, \quad (a^x)^y = a^{x\cdot y}.$$

EXPONENTIAL FUNCTION
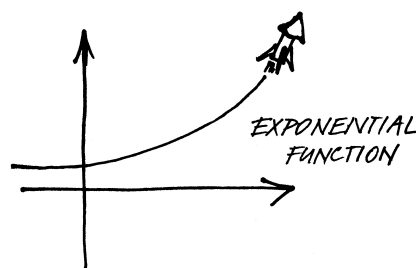
5.2.23. Logarithmic functions. The exponential function $f(x) = a^x$ is increasing for $a > 1$ and decreasing for $0 < a < 1$. Thus in both cases, there is a function $f^{-1}(x)$ inverse to it. This function is called the *logarithmic function with base a*. We write $\ln_a(x)$. $\ln_a(a^x) = x$ is the defining property.

The equalities (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.

we obtain the following formula for approximating the function value by its differential:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

So, setting

$$f(x) := \operatorname{arccotg} x, \quad x_0 := 1,$$

we get

$$\operatorname{arccotg} 1.02 \approx \operatorname{arccotg} 1 + \tfrac{-1}{1+1^2}(1.02 - 1) = \tfrac{\pi}{4} - 0.01.$$

Eventually, let us remark that the point x_0 is of course selected so that the expression $x - x_0$ is as close to zero as possible, yet we must be able to calculate the values of f and f' at the point. □

5.E.7. Using differential, approximate $\arcsin 0.497$. ○

5.E.8. Using differential, approximate $a = \arctan 1.02$ and $b = \sqrt[3]{70}$. ○

5.E.9. Using differential, approximate a) $\sin\left(\tfrac{29\pi}{180}\right)$, b) $\sin\left(\tfrac{46\pi}{180}\right)$. ○

5.E.10. For which $a \in \mathbb{R}$ is the cubic polynomial P which satisfies the conditions $P(0) = 1$, $P'(0) = 1$, $P(1) = 2a+2$, $P'(1) = 5a + 1$, a monotonic function on the whole $\mathbb{R}$?

Solution. From the conditions $P(0) = 1$ and $P'(0) = 1$ it follows that $P(x) = bx^3 + cx^2 + x + 1$ where $b, c \in \mathbb{R}$; 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 $b = c = a$. The polynomials which satisfy the desired conditions are thus of the form $P(x) = ax^3 + ax^2 + x + 1$, $a \in \mathbb{R}$. The monotonicity of the polynomial is equivalent to having no local extrema. The extrema can occur only at those points where the derivative changes sign. Therefore, the polynomial is monotonic if and only if its derivative keeps the sign on the whole $\mathbb{R}$. The derivative is

$$P'(x) = 3ax^2 + 2ax + 1$$

and it will keep the sign iff the discriminant is non-positive. Thus we get the condition

$$\begin{aligned} 4a^2 - 12a &\leq 0 \\ 4a(a - 3) &\leq 0, \end{aligned}$$

which is true for $a \in [0, 3]$. However, for $a = 0$ the polynomial P is monotonic, yet not cubic, Thus the set of satisfactory numbers a is the interval $(0, 3]$. □

There is an extremely important value of a, namely the number e, see the paragraph 5.4.1, sometimes 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 δx 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" δx.

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

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = a$$

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.
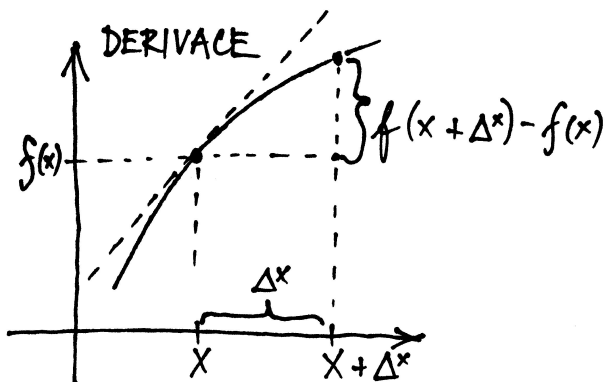
Obviously, the derivative of a complex valued function $(f(x) + i\,g(x))$ exists if and only the derivatives of both real and imaginary parts f and g exist (see the elementary properties of limits). Then

$$(f(x) + i\,g(x))' = f'(x) + i\,g'(x).$$

5.E.11. Determine the parameter $c \in \mathbb{R}$ so that the tangent line to the graph of the function $\frac{\ln(c \cdot x)}{\sqrt{x}}$ at the point $[1, 0]$ goes through the point $[2, 2]$.

Solution. From the statement of the problem it follows that the tangent's slope is $\frac{2-0}{2-1} = 2$. The slope is determined by the derivative at the given point, thus we get the condition

$$\left.\frac{2 - \ln(cx)}{2x\sqrt{x}}\right|_{x=1} = 2, \text{ that is } \quad 2 - \ln(c) = 4,$$

hence $c = \frac{1}{e^2}$. Yet for $c = \frac{1}{e^2}$, the function $\frac{ln(c \cdot x)}{\sqrt{x}}$ takes the value -2 at the point 1. Therefore, there is no such c. $\qquad \square$

5.E.12. A highway patrol helicopter is flying 3 kilometers above a highway at the speed of 120 kph. Its pilot localizes a car whose straight-line distance from the helicopter is 5 kilometers. The car goes in the opposite direction and approaches the helicopter at 160 kph (with regard to the helicopter). Determine the car's speed with regard to a tin lying on the highway.

Solution. For the sake of simplicity, we will omit units of measurement (distances will be expressed in kilometers and times in hours, speeds in kph, then). The helicopter's position at time t can be expressed by the point $[y(t), 3]$, and the car's position by $[x(t), 0]$, then. (We choose the axes so that the helicopter and the car are moving along the x-axis.) Let us denote by $s(t)$ the straight-line distance of the car from the helicopter and by t_0 the moment mentioned in the problem's statement. Let us calculate the car's speed with respect to the origin. We can suppose that $x(t) > y(t) > 0$, then $x'(t) \leq 0$, $y'(t) \geq 0$ for the considered time moments t since the car is approaching the point $[0, 0]$ from the right – the value $x(t)$ decreases as t increases, therefore $x'(t) \leq 0$. Similarly we can get that $y'(t) \geq 0$ and also $s'(t) \leq 0$. Let us add that, for instance, $y'(t)$ determines the rate of change of the function y at time t, i. e. the helicopter's speed

We know that

$$s(t_0) = 5, \quad s'(t_0) = -160, \quad y'(t_0) = 120$$

and that ($s(t)$ is the hypotenuse of the right triangle)

$$(1) \qquad (x(t) - y(t))^2 + 3^2 = s^2(t).$$

Hence it follows ($x(t) > y(t) > 0$) that

$$(x(t_0) - y(t_0))^2 + 3^2 = 5^2, \quad \text{i. e.} \quad x(t_0) - y(t_0) = 4.$$

By differentiating the identity (1), we get

$$2(x(t) - y(t))(x'(t) - y'(t)) = 2s(t)s'(t)$$

and then for $t = t_0$,

Derivatives are handled rather easily, but it takes time to derive the proper formulae for derivatives of the elementary functions in the zoo. Therefore, we present a table of derivatives of several such functions in advance. In the last column, there are references to the corresponding paragraph where the result is 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 $\mathbb{R}$	only the first derivative $h'(x)$ is continuous	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 $\mathbb{R}$	$f'(x) = \ln(a) \cdot a^x$	5.3.7
logarithmic function $f(x) = \ln_a(x)$, $a > 0, a \neq 1$	interval $(0, \infty)$	$f'(x) = (\ln(a))^{-1} \cdot \frac{1}{x}$	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 ψ 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 ψ exists, it is of the form

$$\psi(x) = \frac{(f(x) - f(x_0))}{x - x_0}$$

$2 \cdot 4 \left(x'\left(t_0 \right) - 120 \right) = 2 \cdot 5 \cdot (-160)$,   i. e. $x'\left(t_0 \right) = -80$.

We have calculated that the car is approaching the tin at 80 kph. It suffices to realize with which units of measurement we worked. Having obtained a negative value is caused by our choice of the coordinate system.   □

5.E.13. A 13 feet long ladder is leaned against a house. Suddenly the base of the ladder slips off and the ladder begins to go down (still touching the house at its other end). When the base of the ladder is 12 feet from the house, it is moving at 5 feet per second from it. At this moment:

(a) What is the speed of the top of the ladder?
(b) What is the rate of change of the area of the triangle delimited by the house, the ladder, and ground?
(c) What is the rate of change of the angle enclosed by the ladder and the ground?

○

5.E.14. Determine whether there is a point in the interval $(0, 4)$ such that the tangent line at this point to the polynomial $x \left(x - 4 \right)^5$ is parallel to the x-axis.   ○

5.E.15. Let $p \in (0, +\infty)$. Write the equation of the tangent to the parabola $2py = x^2$ at a general point $[x_0, ?]$.   ○

5.E.16. Find the equation of the normal line to the graph of the function $y = 1 - e^{\frac{x}{2}}$, $x \in \mathbb{R}$ at the point where the graph intersects the x-axis.   ○

5.E.17. Find the equations of the tangent and normal lines to the curve $y = \left(x + 1 \right) \sqrt[3]{3 - x}$, $x \in \mathbb{R}$, at the point $[-1, 0]$.
○

5.E.18. Let the function $y = \frac{\ln\left(2x^3 + 4x^2 - x\right)}{1 + x}$, $x \in \left(\frac{1}{2}, +\infty \right)$ be given. Determine the equations of the tangent and normal lines to the graph of this function at the point $[1, ?]$.   ○

5.E.19. At which points is the tangent to the parabola $y = 2 + x - x^2$, $x \in \mathbb{R}$, parallel to the x-axis?   ○

5.E.20. Determine the equations of the tangent line t and the normal line n to the graph of the function

$$y = \sqrt{x^2 - 3x + 11}, \quad x \in \mathbb{R}$$

at the point $[2, ?]$. Further, determine all points at which the tangent is parallel to the x-axis.   ○

5.E.21. What is the angle between the x-axis and the graph of the function $y = \ln x$? (We mean the angle between the

for all $x \in \mathcal{O} \setminus \{x_0\}$.

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,

$$\lim_{x \to x_0} \psi(x) = f'(x_0) = \psi(x_0).$$

Thus ψ is continuous at x_0 as desired.

On the other hand if such a function ψ 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.   □

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 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 is increasing at x_0. By the previous lemma, $f(x) = f(x_0) + \psi(x)(x - x_0)$ and $\psi(x_0) > 0$. Since ψ 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.   □

A 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 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 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.

tangent line and the positive x-axis in the "positive sense of rotation", ie. counterclockwise.) ○

5.E.22. Determine the equations of the tangent and normal line to the curve given by the equation $x^3 + y^3 - 2xy = 0$ at the point $[1, 1]$. ○

5.E.23. Prove: $\frac{x}{1+x} < \ln(1 + x) < x$ for all $x > 0$. ○

F. Extremal problems

The simple observation 5.3.2 about the geometrical meaning of the derivative also tells us that a differentiable real-valued function of a real variable can have extremes only at the points where its derivative is zero. We can utilize this mere fact when solving miscellaneous practical problems.

5.F.1. Consider the parabola $y = x^2$. Determine the x-coordinate x_A of the parabola's point which is nearest to the point $A = [1, 2]$.

Solution. It is not difficult to realize that there is a unique solution to this problem and that we are actually looking for the absolute minimum of the function
$$f(x) = \sqrt{(x-1)^2 + (x^2 - 2)^2}, \quad x \in \mathbb{R}.$$
Since taking the square root is an increasing function, the function f takes the least value at the same point where the function
$$g(x) = (x-1)^2 + (x^2 - 2)^2, \quad x \in \mathbb{R}$$
does. Since
$$g'(x) = 4x^3 - 6x - 2, \quad x \in \mathbb{R},$$
by solving the equation $0 = 2x^3 - 3x - 1$, we first get the stationary point $x = -1$ and after dividing the polynomial $2x^3 - 3x - 1$ by the polynomial $x + 1$, we then obtain the remaining two stationary points
$$\frac{1-\sqrt{3}}{2} \quad \text{and} \quad \frac{1+\sqrt{3}}{2}.$$
As the function g is a polynomial (differentiable on the whole domain), from the geometrical sense of the problem, we get
$$x_A = \tfrac{1+\sqrt{3}}{2}.$$

□

5.F.2. 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.

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 δ-neighbourhood of y on which $z < y$ implies $f(z) < f(y)$. Let δ_0 be the infimum 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 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. 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
$$f(x) = x + 5x^2 \sin(1/x), \ f(0) = 0.$$
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.
$$\lim_{x \to 0} \frac{x + 5x^2 \sin(1/x)}{x} = \lim_{x \to 0} (1 + 5x \sin(1/x)) = 1.$$
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. The one-sided derivatives of this function do not exist at zero either.
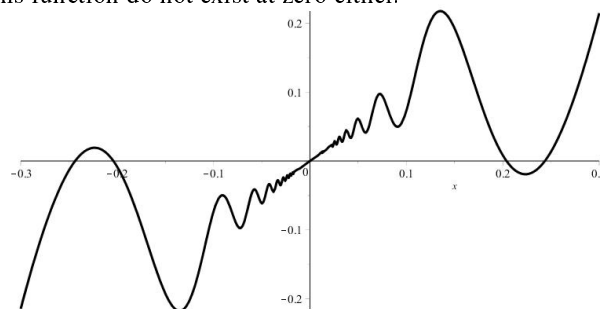
As another illustration of a simple usage of the relation between the derivatives and the properties of being an increasing (or decreasing) function, we can consider the existence of inverses to polynomials.

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

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

$$f(x) = x\left(h - \tfrac{2hx}{b}\right)$$

on the interval $I = [0, b/2]$. Since we have that

$$f'(x) = h - \tfrac{4hx}{b} \quad \text{for all } x \in I$$

and further

$$f(0) = f\left(\tfrac{b}{2}\right) = 0, \qquad f(x) \geq 0, \quad x \in I,$$

the function f must take the greatest value on I at its only stationary point $x_0 = b/4$. Thus the sides of the wanted rectangle are $b/2$ long (twice x_0: considering the original problem) and $h/2$ (which can be obtained by substituting $b/4$ for x into the expression $h - 2hx/b$). Hence we get $S = hb/4$. $\qquad\square$

5.F.3. Among rectangles such that two of their vertices lie on the x-axis and the other two have positive y-coordinates and lie on the parabola $y = 8 - 2x^2$, find the one which has the greatest area.

Solution. The base of the largest rectangle is $4/\sqrt{3}$ long, the rectangle's height is then $16/3$. This result can be obtained by finding the absolute maximum of the function

$$S(x) = 2x\left(8 - 2x^2\right)$$

on the interval $I = [0, 2]$. Since this function is non-negative on I, takes zero at I's boundary points, is differentiable on the whole of I and its derivative is zero at a unique point of I, namely $x = 2/\sqrt{3}$, it has the maximum there. $\qquad\square$

5.F.4. Let the ellipse $3x^2 + y^2 = 2$ be given. Write the equation of its tangent line which forms the smallest triangle possible in the first quadrant and determine the triangle's area.

Solution. The line corresponding to the equation $ax + by + c = 0$ intersects the axes at the points $[-\tfrac{c}{a}, 0]$, $[0, -\tfrac{c}{b}]$ and the area of the triangle whose vertices are these two points and the origin is $S = \tfrac{c^2}{2ab}$. The line which touches the ellipse at $[x_T, y_T]$ has the equation $3xx_T + yy_T - 2 = 0$. The area of the triangle corresponding to it is thus $S = \tfrac{2}{3x_T y_T}$. Further, in the first quadrant, we have that $x_T, y_T > 0$. To minimize this area means to maximize the product $x_T y_T = x_T\sqrt{2 - 3x_T^2}$, which is (in the first quadrant) the same as to maximize $(x_T y_T)^2 = x_T^2(2 - 3x_T^2) = -3(x_T^2 - \tfrac{1}{3})^2 + \tfrac{1}{3}$. Hence, the wanted minimum is at $x_T = \tfrac{1}{\sqrt{3}}$. The tangent's equation is $\sqrt{3}x + y = 2$ and the triangle's area is $S_{min} = \tfrac{2\sqrt{3}}{9}$. $\qquad\square$

them. On the other hand, the inverse exists to every restriction 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 $(-\infty, -\tfrac{b}{2a})$, $(-\tfrac{b}{2a}, \infty)$.

It can be shown that the roots of polynomials of order larger than four cannot in general be expressed by means of power functions. Thus piece-wise defined inverses to polynomials may represent new items in the 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

$$\delta y = f(x + \delta x)g(x + \delta x) - f(x)g(x)$$
$$= f(x + \delta x)(g(x + \delta x) - g(x)) + (f(x + \delta x) - f(x))g(x)$$

Now, if we make the increase δx 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 $fg' + f'g$, which is called the *Leibniz rule*[10].

---

[10] 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.

5.F.5. At the time $t = 0$, the three points P, Q, R began moving in the plane as follows: The point P is moving from the point $[-2, 1]$ in the direction $(3, 1)$ at the constant speed $\sqrt{10}$ m/s. the point Q is moving from $[0, 0]$ in the direction $(-1, 1)$ with the constant acceleration $2\sqrt{2}$ m/s^2 (beginning at zero speed) and the point R is going from $[0, 1]$ in the direction $(1, 0)$ at the constant speed 2 m/s. At which time will the area of the triangle PQR be minimal?

Solution. The equations of the points P, Q, R in time are

$$
\begin{aligned}
P &: \quad [-2, 1] + (3, 1)t, \\
Q &: \quad [0, 0] + (-1, 1)t^2, \\
R &: \quad [0, 1] + (2, 0)t.
\end{aligned}
$$

The area of the triangle PQR is determined, for instance, by half the absolute value of the determinant whose rows are the coordinates of the vectors PQ and QR (see 1.5.11). So we minimize the determinant:

$$
\begin{vmatrix} -2 + t & t \\ -t^2 - 2t & -1 + t^2 \end{vmatrix} = 2t^3 - t + 2.
$$

The derivative is $6t^2 - 1$, so the extrema occur at $t = \pm\frac{1}{\sqrt{6}}$. Thanks to considering non-negative time only, we are interested in $t = \frac{1}{\sqrt{6}}$. The second derivative of the considered function is positive at this point, thus the function has a local minimum there. Further, its value at this point is positive and less than the value at the point 0 (the boundary point of the interval where we are looking for the extremum), so this point is the wanted global minimum. $\square$

5.F.6. At 9 o'clock in the morning, the old wolf left his den D and as a part of his everyday warm-up, he began running counterclockwise around his favorite stump S at the constant speed 4 kph (not very quick, is he), keeping the constant distance of 1 km from it. At the same time, Little Red Riding Hood set out from her house H straight to her Grandma's cottage C at the constant speed 4 kph. When will they be closest to each other and what will their distance be at that time? The coordinates (in kilometers) are: $D = [2, 3]$, $S = [2, 2]$, $H = [0, 0]$, $C = [5, 5]$.

Solution. The wolf is moving along a unit circle, so his angular speed equals his absolute speed and his position in time can be described by the following parametric equations:

$$
x(t) = 2 - \cos(4t), \quad y(t) = 2 - \sin(4t),
$$

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 $x \to x_0$ gives the desired result because f is continuous at x_0.

(4) By lemma 5.3.1, there are functions ψ and φ 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

Little Red Riding Hood is then moving along the line

$$x(t) = 2\sqrt{2}t, \ y(t) = 2\sqrt{2}t.$$

Let us find the extrema of the (squared) distance ρ of their paths in time:

$$\rho(t) = [2 - \cos(4t) - 2\sqrt{2}t]^2 + [2 - \sin(4t) - 2\sqrt{2}t]^2,$$
$$\rho'(t) = 16(\cos(4t) - \sin(4t))(\sqrt{2}t - 1) + 32t +$$
$$+ 4\sqrt{2}(\cos(4t) + \sin(4t)) - 16\sqrt{2}.$$

It is impossible to solve the equation $\rho'(t) = 0$ algebraically, we can only find the solution numerically (using some computational software). Apparently, there will be infinitely many extrema: every round, the wolf's direction is at some moment parallel to that of Little Red Riding Hood, so their distance is decreasing for some period; however, Little Red Riding Hood is moving away from the wolf's favorite stump around which he is moving. We find out that the first local minimum occurs at $t \doteq 0.31$, and then at $t \doteq 0.97$, when the distance of our heroes will be approximately 5 meters. Clearly this is the global maximum as well.

The situation when we cannot solve a given problem explicitly is quite common in practice and the use of numerical methods is of great importance. □

5.F.7. Halley's problem, 1686. A basketball player is standing in front of a basket, at distance l from its rim which is at height h from the throwing point. Determine the minimal initial speed v_0 which the player must give to the ball in order to score, and the angle φ corresponding to this v_0. See the picture.

Solution. Once again, we will omit units of measurement: we can assume that distances are given in meters, times in seconds (and speeds in meters per second then). Suppose the player throws the ball at time $t = 0$ and it goes through the rim at time $t_0 > 0$. We will express the ball's position (while flying) by the points $[x(t), y(t)]$ for $t \in [0, t_0]$, and we require that $x(0) = 0$, $y(0) = 0$, $x(t_0) = l$, $y(t_0) = h$.

Apparently,

$$x'(t) = v_0 \cos\varphi, \qquad y'(t) = v_0 \sin\varphi - gt$$

for $t \in (0, t_0)$, where g is the gravity of Earth, since the values $x'(t)$ and $y'(t)$ are, respectively, the horizontal and vertical speed of the ball. By integrating these equations, we get

$$x(t) = v_0 t \cos\varphi + c_1, \qquad y(t) = v_0 t \sin\varphi - \tfrac{1}{2}gt^2 + c_2$$

for $t \in (0, t_0)$ and $c_1, c_2 \in \mathbb{R}$. From the initial conditions

its value at x_0 is just the desired derivative of the composite function, again by lemma 5.3.1. □

5.3.5. Derivative of quotients. Consider first the special case of $h(x) = x^{-1}$. From the definition of the derivative,

$$h'(x) = \lim_{\delta x \to 0} \frac{\frac{1}{x + \delta x} - \frac{1}{x}}{\delta x} = \lim_{\delta x \to 0} \frac{x - x - \delta x}{\delta x(x^2 + x\delta x)}$$
$$= \lim_{\delta x \to 0} \frac{-1}{x^2 + x\delta x} = -x^{-2}.$$

Thus, the above leads to:

<div align="center">

DERIVATIVE OF A QUOTIENT

</div>

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

$$(f/g)' = (f \cdot g^{-1})' = f'g^{-1} - fg^{-2}g' = \frac{f'g - gf'}{g^2}.$$

□

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 $f : \mathbb{R} \to \mathbb{R}$ 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

$$f^{-1} \circ f = \mathrm{id}_\mathbb{R}, \quad f \circ f^{-1} = \mathrm{id}_\mathbb{R},$$

. 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 id_A and 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.

INVERSION FUNCTION
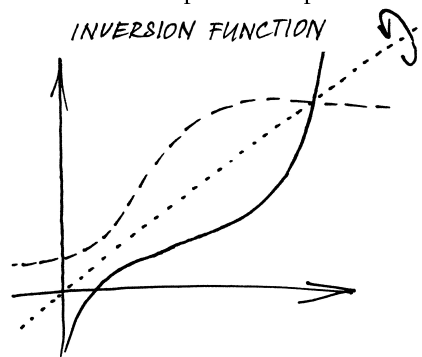
$$\lim_{t\to 0+} x(t) = x(0) = 0, \qquad \lim_{t\to 0+} y(t) = y(0) = 0,$$

it follows that $c_1 = c_2 = 0$. Substituting the remaining conditions

$$\lim_{t\to t_0-} x(t) = x(t_0) = l, \qquad \lim_{t\to t_0-} y(t) = y(t_0) = h$$

then gives

$$l = v_0 t_0 \cos\varphi, \qquad h = v_0 t_0 \sin\varphi - \tfrac{1}{2}\,g t_0^2.$$

According to the first equation, we have that

$$(1) \qquad\qquad t_0 = \frac{l}{v_0 \cos\varphi},$$

and thus we get only one equation

$$(2) \qquad\qquad h = l\tan\varphi - \frac{g l^2}{2 v_0^2 \cos^2\varphi},$$

where $v_0 \in (0, +\infty)$, $\varphi \in (0, \pi/2)$.

Let us remind that our task is to determine the minimal v_0 and the corresponding φ which satisfies this equation. To be more comprehensible, we want to find the minimal value of v_0 for which there is an angle φ satisfying (2). Since

$$\frac{1}{\cos^2\varphi} = \frac{\cos^2\varphi + \sin^2\varphi}{\cos^2\varphi} = 1 + \tan^2\varphi, \qquad \varphi \in \left(0, \tfrac{\pi}{2}\right),$$

the equation (2) can be written in the form

$$h - l\tan\varphi + \frac{g l^2}{2 v_0^2}\left(1 + \tan^2\varphi\right) = 0,$$

i. e.

$$\tan^2\varphi - \frac{2 v_0^2}{g l}\tan\varphi + \frac{2 h v_0^2}{g l^2} + 1 = 0.$$

From the last equation (quadratic equation in $p = \tan\varphi$), it follows that

$$\tan\varphi = \frac{\frac{2 v_0^2}{g l} \pm \sqrt{\frac{4 v_0^4}{g^2 l^2} - 4\left(\frac{2 h v_0^2}{g l^2} + 1\right)}}{2},$$

i. e.

$$(3) \qquad \tan\varphi = \frac{v_0^2}{g l} \pm \frac{\sqrt{v_0^4 - 2 h v_0^2 g - g^2 l^2}}{g l}.$$

Therefore, the angle φ satisfying (2) exists if and only if

$$v_0^4 - 2 g h\, v_0^2 - g^2 l^2 \ge 0.$$

Once again, a suitable substitution (this time $q = v_0^2$) allows us to reduce the left side to a quadratic expression and subsequently to get

$$\left(v_0^2 - g\left[h + \sqrt{h^2 + l^2}\right]\right)\left(v_0^2 - g\left[h - \sqrt{h^2 + l^2}\right]\right) \ge 0.$$

As $h < \sqrt{h^2 + l^2}$, it must be that

$$v_0^2 \ge g\left[h + \sqrt{h^2 + l^2}\right], \quad \text{i. e.} \quad v_0 \ge \sqrt{g\left[h + \sqrt{h^2 + l^2}\right]}.$$

The least value

$$(4) \qquad\qquad v_0 = \sqrt{g\left[h + \sqrt{h^2 + l^2}\right]}$$

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

$$1 = (\mathrm{id})'(x) = (f \circ f^{-1})'(x) = f'(y) \cdot (f^{-1})'(x).$$

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 $f'(y_0) \ne 0$, then*

$$(1) \qquad (f^{-1})'(x_0) = \frac{1}{f'(f^{-1}(x_0))} = \frac{1}{f'(y_0)}.$$

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 $f \circ f^{-1} = \mathrm{id}$ there. The composite function is differentiable. By lemma 5.3.1, there is a function φ 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 $\varphi(y_0) = f'(y_0) \ne 0$ and φ 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 ε and δ 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

$$\frac{f^{-1}(x) - f^{-1}(x_0)}{x - x_0} = \frac{1}{\varphi(f^{-1}(x))} \ne 0,$$

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

$$\lim_{x\to x_0} \frac{1}{\varphi(f^{-1}(x))} = \frac{1}{\varphi(f^{-1}(x_0))} = \frac{1}{f'(y_0)}.$$

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. $\square$

is then matched (see (3)) by

$$\tan \varphi = \frac{v_0^2}{gl} = \frac{h + \sqrt{h^2 + l^2}}{l},$$

(5)

$$\text{i. e.} \quad \varphi = \operatorname{arctg} \frac{h + \sqrt{h^2 + l^2}}{l}.$$

The previous calculation was based upon the conditions $x(t_0) = l$, $y(t_0) = h$ only. However, these only talk about the position of the ball at the time t_0, but the ball could get through the rim from below. Therefore, let us add the condition $y'(t_0) < 0$ which says that the ball was falling at the time, and let us prove that it holds for v_0 from (4) and φ from (5).

Let us remind that we have (see (1), (2))

$$t_0 = \frac{l}{v_0 \cos \varphi}, \qquad v_0^2 = \frac{gl^2}{2(l \tan \varphi - h) \cos^2 \varphi}.$$

Using this, from

$$y'(t_0) = \lim_{t \to t_0-} y'(t) = v_0 \sin \varphi - gt_0 < 0$$

we get

$$\frac{gl^2}{2(l \tan \varphi - h) \cos^2 \varphi} = v_0^2 < v_0 \cdot \frac{gt_0}{\sin \varphi} = \frac{gl}{\sin \varphi \cos \varphi},$$

i. e. the equality

$$l \sin \varphi \cos \varphi < 2(l \tan \varphi - h) \cos^2 \varphi,$$

from which we can easily see that

$$\frac{2h}{l} < \tan \varphi.$$

By confrontation with (5), we get that the last inequality really holds because

$$\tan \varphi = \frac{h + \sqrt{h^2 + l^2}}{l} > \frac{h + \sqrt{h^2}}{l} = \frac{2h}{l}.$$

Thus we have shown that for the initial speed from (4), the player is able to score.

During the free throw, supposing the player lets the ball go at the height of 2 m, we have

$$h = 1.05 \,\text{m}, \quad l = 4.225 \,\text{m}, \quad g = 9.806\,65 \,\text{m} \cdot \text{s}^{-2},$$

and so the minimal initial speed of the ball is

$$v_0 = \sqrt{9.806\,65 \left[1.05 + \sqrt{(1.05)^2 + (4.225)^2} \right]} \,\text{m} \cdot \text{s}^{-1} \doteq$$
$$7.28 \,\text{m} \cdot \text{s}^{-1}.$$

The corresponding angle is then

$$\varphi = \operatorname{arctg} \frac{v_0^2}{9.806\,65 \cdot 4.225} \doteq 0.907 \,\text{rad} \approx 52\,°.$$

Let us think for a while about the obtained value of the angle φ for the initial speed v_0. According to the picture, we have

$$2\beta + (\pi - \alpha) = \pi \qquad \text{and} \qquad \alpha + \gamma = \tfrac{\pi}{2},$$

whence it follows that

$$\beta = \tfrac{\alpha}{2} = \tfrac{\pi}{4} - \tfrac{\gamma}{2}.$$

So it holds that

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

$$f'(x) = \lim_{\delta x \to 0} \frac{a^{x + \delta x} - a^x}{\delta x} = a^x \lim_{\delta x \to 0} \frac{a^{\delta x} - 1}{\delta x} = f'(0)a^x.$$

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, i, and 6.3.7).

There is an especially important base e, sometimes 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) $$(\ln)'(y) = (\ln)'(e^x) = \frac{1}{(e^x)'} = \frac{1}{e^x} = \frac{1}{y}.$$

The formula

(2) $$(x^a)' = ax^{a-1}$$

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.

$\varphi = \frac{\pi}{2} - \beta = \frac{\pi}{4} + \frac{\gamma}{2} = \frac{1}{2}\left(\frac{\pi}{2} + \gamma\right) = \frac{1}{2}\left(\frac{\pi}{2} + \operatorname{arctg} \frac{h}{l}\right).$

We have obtained that the elevation angle corresponding to the throw with minimal energy is the arithmetic mean of the right angle and the angle at which the rim is seen (from the ball's position).

The problem of finding the minimal speed of the thrown ball was actually solved by Edmond Halley as early as in 1686, when he determined the minimal amount of gunpowder necessary for a cannonball to hit a target which lies at greater height (beyond a rampart, for instance). Halley proved (the so-called Halley's calibration rule) that to hit a target at the point $[l, h]$ (shooting from $[0, 0]$) one needs the same minimal amount of gunpowder as when hitting a horizontal target at distance $h + \sqrt{h^2 + l^2}$ (at the angle $\varphi = 45\,^\circ$). Halley also demonstrated that the value of φ is stable with regard to small difference of the amount of used gunpowder and insignificant errors in estimating the target's distance. □

5.F.8. A bullet is shot at angle φ from a point at height h above ground at initial speed v_0. It will fall on the ground at distance R from the point of shot (see the picture). Determine the angle φ for which the value of R is maximal.

Solution. We will express the bullet's position in time by the points $[x(t), y(t)]$. We assume that it was shot at time $t = 0$ from the point $[0, 0]$ and it will fall on the ground at the point $[R, -h]$ at certain time $t = t_0$, i. e. $x(0) = 0, y(0) = 0,$ $x(t_0) = R, y(t_0) = -h$. Similarly to Halley's problem, we will consider the equations

$$x'(t) = v_0 \cos \varphi, \quad y'(t) = v_0 \sin \varphi - gt, \qquad t \in (0, t_0)$$

for the horizontal and vertical speeds of the bullet, where g is the gravity of Earth.

We can continue as when solving the previous problem: by integrating these equations (taking $x(0) = y(0) = 0$ into consideration), we get

ROLLE'S THEOREM[11]

Theorem. *Assume that the function* $f : \mathbb{R} \to \mathbb{R}$ *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* $f'(c) = 0$.
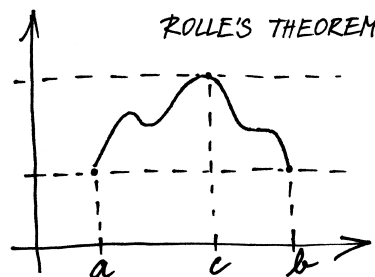
PROOF. Since the function f is continuous on the closed interval (i.e. on a compact set), it attains its maximum and its minimum there. Either its maximum value is greater than $f(a) = f(b)$, or the minimum value is less than $f(a) = f(b)$, or f is constant. If the third case applies, the derivative 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. □

5.3.9. The latter result immediately implies the following corollary.

LAGRANGE'S MEAN VALUE THEOREM

Theorem. *Assume the function* $f : \mathbb{R} \to \mathbb{R}$ *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*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

---
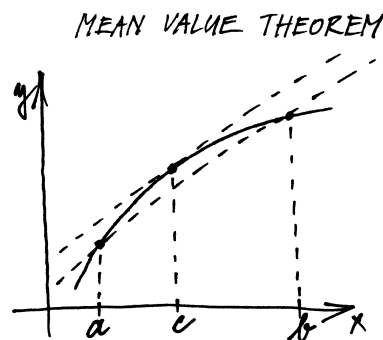
[11] The French mathematician Michel Rolle (1652-1719) proved this theorem only for polynomials. The principle was perhaps known much earlier, but the rigorous proof comes from the 19th century only.

$$x(t) = v_0 t \cos\varphi, \quad y(t) = v_0 t \sin\varphi - \tfrac{1}{2} g t^2, \qquad t \in (0, t_0),$$

and from the conditions $\lim_{t \to t_0-} x(t) = x(t_0) = R$, $\lim_{t \to t_0-} y(t) = y(t_0) = -h$, we then have that

$$R = v_0 t_0 \cos\varphi, \quad -h = v_0 t_0 \sin\varphi - \tfrac{1}{2} g t_0^2.$$

From the first equation, it follows that

$$t_0 = \tfrac{R}{v_0 \cos\varphi},$$

so we can express the previous two equations by the single equation

$$(1) \qquad -h = R \tan\varphi - \frac{gR^2}{2 v_0^2 \cos^2\varphi},$$

where $\varphi \in (0, \pi/2)$.

Unlike with Halley's problem, the value of v_0 is given and R is variable (dependent on φ). So, actually, there is a function $R = R(\varphi)$ (in variable φ) which must satisfy (1) (it is determined by the equation (1)). Thus, this function is given implicitly. The equation (1) can be written as (R is substituted by $R(\varphi)$)

$$R(\varphi) \tan\varphi \cdot 2 v_0^2 \cos^2\varphi - gR^2(\varphi) + h \cdot 2 v_0^2 \cos^2\varphi = 0.$$

Using the relation

$$2 \tan\varphi \cos^2\varphi = \sin 2\varphi,$$

we can transform (1) into the form

$$(2) \quad R(\varphi) v_0^2 \sin 2\varphi - gR^2(\varphi) + 2h v_0^2 \cos^2\varphi = 0.$$

Differentiating with respect to φ now gives

$$R'(\varphi) v_0^2 \sin 2\varphi + 2R(\varphi) v_0^2 \cos 2\varphi - 2gR(\varphi)R'(\varphi) - 2h v_0^2 (2\cos\varphi \sin\varphi) = 0,$$

i. e.

$$R'(\varphi) \left[v_0^2 \sin 2\varphi - 2gR(\varphi) \right] = -2R(\varphi) v_0^2 \cos 2\varphi + 2h v_0^2 \sin 2\varphi.$$

Thus we have calculated that

$$R'(\varphi) = \tfrac{2 v_0^2 [h \sin 2\varphi - R(\varphi) \cos 2\varphi]}{v_0^2 \sin 2\varphi - 2gR(\varphi)}, \qquad \varphi \in \left(0, \tfrac{\pi}{2}\right).$$

It suffices to verify that $v_0^2 \sin 2\varphi - 2gR(\varphi) \neq 0$ for every $\varphi \in (0, \pi/2)$. Let us suppose the contrary and substitute

$$R = \tfrac{v_0^2 \sin 2\varphi}{2g} = \tfrac{v_0^2 \sin\varphi \cos\varphi}{g}$$

into (1), obtaining

$$-h = \tfrac{v_0^2 \sin\varphi \cos\varphi}{g} \tan\varphi - \tfrac{g v_0^4 \sin^2\varphi \cos^2\varphi}{2g^2 v_0^2 \cos^2\varphi}.$$

Simple rearrangements lead to

$$-h = \tfrac{v_0^2 \sin^2\varphi}{2g},$$

which cannot happen (the left side is surely negative while the right one is positive).

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 $h(a) = h(b)$ and

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

By the previous theorem, there is a point c at which $h'(c) = 0$. $\qquad \square$

The mean value theorem can also be written in the form:

$$(1) \qquad f(b) = f(a) + f'(c)(b - a).$$

In the case of a parametrically given curve in the plane, i.e. a pair of functions $y = f(t)$, $x = g(t)$, the same 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*:

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, and further let $g'(t) \neq 0$ for all $t \in (a, b)$. 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 $h(a) = f(b)g(a) - f(a)g(b)$, $h(b) = f(b)g(a) - f(a)g(b)$, so by Rolle's theorem, there is a number $c \in (a, b)$ such that $h'(c) = 0$.

Finally, the function g is either strictly increasing or decreasing on $[a, b]$ and thus $g(b) \neq g(a)$. Moreover, $g'(c) \neq 0$ and the desired formula follows. $\qquad \square$

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.

So we were able to determine $R'(\varphi)$ for all $\varphi \in (0, \pi/2)$. What is more, we can immediately see that this derivative is zero if and only if

$$h \sin 2\varphi = R(\varphi) \cos 2\varphi, \qquad \text{i. e.} \qquad R(\varphi) = h \tan 2\varphi.$$

Since the function R must have a maximum on the interval $(0, \pi/2)$ (according to the physical meaning of the problem, for $\varphi \to 0+$ and for $\varphi \to \pi/2-$ the value of R decreases) and is differentiable at every point of this interval, it has its maximum at the point where its derivative is zero. This means that $R(\varphi)$ can be maximal only if

$$(3) \qquad\qquad R(\varphi) = h \tan 2\varphi.$$

Let us thus substitute (3) into (2). We obtain

$$h \tan 2\varphi \, v_0^2 \sin 2\varphi - gh^2 \tan^2 2\varphi + 2hv_0^2 \cos^2 \varphi = 0,$$

and let us transform this equation:

$$\tan 2\varphi \, v_0^2 \sin 2\varphi + 2v_0^2 \cos^2 \varphi = gh \tan^2 2\varphi,$$
$$v_0^2 \frac{\sin^2 2\varphi}{\cos 2\varphi} + v_0^2 (\cos 2\varphi + 1) = gh \frac{\sin^2 2\varphi}{\cos^2 2\varphi},$$
$$v_0^2 \sin^2 2\varphi + v_0^2 \cos^2 2\varphi + v_0^2 \cos 2\varphi = gh \frac{\sin^2 2\varphi}{\cos 2\varphi}$$
$$v_0^2 + v_0^2 \cos 2\varphi = gh \frac{1 - \cos^2 2\varphi}{\cos 2\varphi},$$
$$v_0^2 (1 + \cos 2\varphi) = gh \frac{(1 - \cos 2\varphi)(1 + \cos 2\varphi)}{\cos 2\varphi},$$
$$v_0^2 \cos 2\varphi = gh (1 - \cos 2\varphi), \text{ and } \cos 2\varphi = \frac{gh}{v_0^2 + gh}.$$

However, by this we have uniquely determined the point

$$\varphi_0 = \tfrac{1}{2} \arccos \frac{gh}{v_0^2 + gh},$$

at which R is highest. Since $\sin 2\varphi_0 = \sqrt{1 - \cos^2 2\varphi_0} =$
$\sqrt{1 - \frac{g^2 h^2}{(v_0^2 + gh)^2}} = \frac{\sqrt{v_0^4 + 2ghv_0^2}}{v_0^2 + gh}$, we have

$$R(\varphi_0) = h \tan 2\varphi_0 = h \frac{\frac{\sqrt{v_0^4 + 2ghv_0^2}}{v_0^2 + gh}}{\frac{gh}{v_0^2 + gh}} = \frac{\sqrt{v_0^4 + 2ghv_0^2}}{g}$$
$$= \frac{v_0}{g} \sqrt{v_0^2 + 2gh}.$$

Let, for instance, javelin thrower Barbora Špotáková give a javelin the speed $v_0 = 27.778$ m/s $\doteq 100$ km/h at the height $h = 1.8$ m (with $g = 9.806\,65$ m $\cdot$ s^{-2}). Then the javelin can fly up to the distance
$R(\varphi_0) = \frac{27.778}{9.806\,65} \sqrt{27.778^2 + 2 \cdot 9.806\,65 \cdot 1.8}$ m $\doteq 80.46$m.

This distance was achieved for
$\varphi_0 = \tfrac{1}{2} \arccos \frac{9.806\,65 \cdot 1.8}{27.778^2 + 9.806\,65 \cdot 1.8} \doteq 0.774\,2$ rad $\approx 44.36°$.

However, the world record of Barbora Špotáková does not even approach 80 m although the impact of other phenomena (air resistance, for example) can be neglected. Still we must not forget that from 1 April 1999, the center of gravity of the women's javelin was moved towards its tip upon the decision of IAAF (International Association of Athletics Federation). This reduced the flight distance by around 10 %.

L'HOPITAL'S RULE[12]

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*

$$\lim_{x \to x_0} f(x) = 0, \qquad \lim_{x \to x_0} g(x) = 0.$$

If the limit

$$\lim_{x \to x_0} \frac{f'(x)}{g'(x)}$$

exists, then the limit

$$\lim_{x \to x_0} \frac{f(x)}{g(x)}$$

also exists, and the two limits are equal.
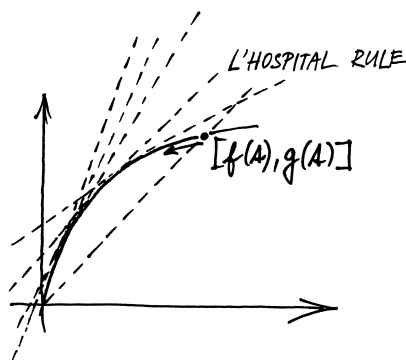
PROOF. Without loss of generality, the functions f and g are zero at the point x_0. The quotient of the values then corresponds to the slope of the secant line between the points $[0, 0]$ and $[f(x), g(x)]$. At the same time, the quotient of the derivatives corresponds to the slope of the tangent line 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, $g'(c) \neq 0$.[13] By the mean value theorem,

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{g(x) - g(x_0)} = \lim_{x \to x_0} \frac{f'(c_x)}{g'(c_x)},$$

---

[12] Guillaume François Antoine, Marquis de l'Hôpital, (1661-1704) became famous for his textbook on Calculus. This rule was first published there, perhaps originally proved by one of the famous Bernoulli brothers.

[13] 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.'

The original record (with "correctly balanced" javelin) was 80.00 m.

The performed reasoning and the obtained result can be applied to other athletic disciplines and sports. In golf, for instance, h is close to 0, and thus it is just the angle

$$\varphi_0 = \lim_{h \to 0+} \tfrac{1}{2} \arccos \frac{gh}{v_0^2 + gh} = \tfrac{1}{2} \arccos 0 = \tfrac{\pi}{4} \text{ rad} = 45°$$

at which the ball falls at the greatest distance

$$R(\varphi_0) = \lim_{h \to 0+} \frac{v_0}{g} \sqrt{v_0^2 + 2gh} = \frac{v_0^2}{g}.$$

Let us realize that our calculation cannot be used for $h = 0$ ($\varphi_0 = \pi/4$) since then we would get the undefined expression $\tan(\pi/2)$ for the distance R. However, we have solved the problem for any $h > 0$, and therefore we could get a helping hand form the corresponding one-sided limit. □

5.F.9. Regiomontanus' problem, 1471.

In the museum, there is a painting on the wall. Its lower

edge is a meters above ground and its upper edge b meters, then (its height thus equals $b - a$). A tourist is looking at the painting, her eyes being at height $h < a$ meters above ground. (The reason for the inequality $h < a$ can, for instance, be to allow more equally tall visitors to view the painting simultaneously in several rows.) How far from the wall should the tourist stand if she wants to maximize her angle of view at the painting?

Solution. Let us denote by x the distance (in meters) of the tourist from the wall and by φ her angle of view at the painting. Further, let us set (see the picture) the angles $\alpha, \beta \in (0, \pi/2)$ by

$$\tan \alpha = \tfrac{b-h}{x}, \qquad \tan \beta = \tfrac{a-h}{x}.$$

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 φ increases as x decreases ($\varphi = \pi$ for $x = 0$ and $h \in (a, b)$).

where c_x is a number lying between x_0 and x, dependent on x. From the existence of the limit

$$\lim_{x \to x_0} \frac{f'(x)}{g'(x)},$$

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 $x_n \to x_0$, and thus the limit

$$\lim_{x \to x_0} \frac{f'(c_x)}{g'(c_x)}$$

exist, and the last two limits are equal. Hence the desired limit exists and has the same value. □

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

$$\lim_{x \to \infty} f(x) = 0, \qquad \lim_{x \to \infty} g(x) = 0,$$

then $\lim_{x \to 0_+} f(1/x) = 0$ and $\lim_{x \to 0_+} g(1/x) = 0$.

At the same time, from existence of the limit of the quotient of the derivatives at infinity,

$$\lim_{x \to 0_+} \frac{(f(1/x))'}{(g(1/x))'} = \lim_{x \to 0_+} \frac{f'(1/x)(-1/x^2)}{g'(1/x)(-1/x^2)}$$
$$= \lim_{x \to 0_+} \frac{f'(1/x)}{g'(1/x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}.$$

Applying the previous theorem, the limit

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to 0_+} \frac{f(1/x)}{g(1/x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}$$

exists in this case as well.

The limit calculation is even simpler in the case when

$$\lim_{x \to x_0} f(x) = \pm\infty, \qquad \lim_{x \to x_0} g(x) = \pm\infty.$$

Then it suffices to write

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{1/g(x)}{1/f(x)},$$

which is already the case of usage of l'Hospital's rule from the previous theorem. It can be proved that l'Hospital's rule has the same form for infinite 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 $\lim_{x \to x_0} f(x) = \pm\infty$ and $\lim_{x \to x_0} g(x) = \pm\infty$ exist. If the limit*
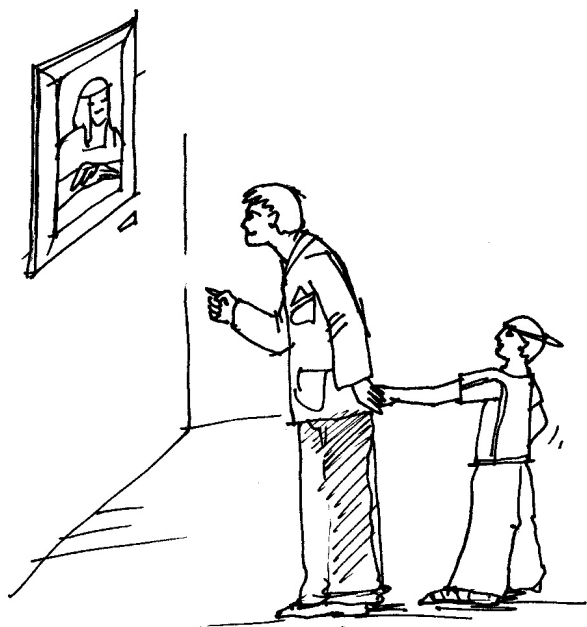
$$\lim_{x \to x_0} \frac{f'(x)}{g'(x)}$$

exists, then the limit

$$\lim_{x \to x_0} \frac{f(x)}{g(x)}$$

also exists and they equal each other.

From the condition $h < a$ it follows that the angle φ is acute, i. e. $\varphi \in (0, \pi/2)$. Since the function $y = \tan x$ is increasing on the interval $(0, \pi/2)$, we can turn our attention to maximizing the value $\tan \varphi$. We have that

$$\tan \varphi = \tan(\alpha - \beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta} = \frac{\frac{b-h}{x} - \frac{a-h}{x}}{1 + \frac{b-h}{x} \cdot \frac{a-h}{x}} = \frac{x(b-a)}{x^2 + (b-h)(a-h)}.$$

So it suffices to find the global maximum of the function

$$f(x) = \frac{x(b-a)}{x^2 + (b-h)(a-h)}, \qquad x \in [0, +\infty).$$

From the expression

$$f'(x) = \frac{(b-a)[x^2 + (b-h)(a-h)] - 2x^2(b-a)}{[x^2 + (b-h)(a-h)]^2} = \frac{(b-a)[(b-h)(a-h) - x^2]}{[x^2 + (b-h)(a-h)]^2}, \quad x \in (0, +\infty),$$

we can see that

$$f'(x) > 0 \quad \text{for} \quad x \in \left(0, \sqrt{(b-h)(a-h)}\right),$$

$$f'(x) < 0 \quad \text{for} \quad x \in \left(\sqrt{(b-h)(a-h)}, +\infty\right).$$

Hence the function f has its global maximum at the point $x_0 = \sqrt{(b-h)(a-h)}$ (let us remind the inequalities $h < a < b$).

The point x_0 can, of course, be determined by other means. For instance, we can (instead of looking for the maximum of the positive function f on the interval $(0, +\infty)$) try to find the global minimum of the function

$$g(x) = \frac{1}{f(x)} = \frac{x^2 + (b-h)(a-h)}{x(b-a)} = \frac{x}{b-a} + \frac{(b-h)(a-h)}{x(b-a)}, \qquad x \in (0, +\infty)$$

with the help of the so-called AM-GM inequality (between the arithmetic and geometric means)

PROOF. Apply the mean value theorem. The key step is to express the quotient in a form where the derivative arises:

$$\frac{f(x)}{g(x)} = \frac{f(x)}{f(x) - f(y)} \cdot \frac{f(x) - f(y)}{g(x) - g(y)} \cdot \frac{g(x) - g(y)}{g(x)},$$

where y is fixed, from a selected 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. As x approaches x_0, the former fraction converges to one. If y is simultaneously moved towards x_0, the latter fraction becomes arbitrarily close to the limit value of the quotient of the derivatives. $\square$

5.3.12. Example. By making suitable modifications of the examined expressions, one can also apply l'Hopital's rule on forms of the types $\infty - \infty$, 1^∞, $0 \cdot \infty$, and so on. Often one simply rearranges the expressions or uses some continuous function, for instance the exponential one.

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*

$$M^1(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$$

is a special case of the *power mean with exponent r*, also known as the generalized mean:

$$M^r(x_1, \ldots, x_n) = \left(\frac{x_1^r + \cdots + x_n^r}{n}\right)^{\frac{1}{r}}.$$

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.

$$\lim_{r \to 0} \ln(M^r(x_1, \ldots, x_n)) = \lim_{r \to 0} \frac{\ln(\frac{1}{n}(x_1^r + \cdots + x_n^r))}{r}$$

$$= \lim_{r \to 0} \frac{\frac{x_1^r \ln x_1 + \cdots + x_n^r \ln x_n}{n}}{\frac{x_1^r + \cdots + x_n^r}{n}}$$

$$= \frac{\ln x_1 + \cdots + \ln x_n}{n} = \ln \sqrt[n]{x_1 \cdots \cdots x_n}.$$

Hence

$$\lim_{r \to 0} M^r(x_1, \ldots, x_n) = \sqrt[n]{x_1 \ldots x_n},$$

324

$$\frac{y_1+y_2}{2} \geq \sqrt{y_1 \, y_2}, \qquad y_1, y_2 \geq 0,$$

where the equality occurs iff $y_1 = y_2$. The choice

$$y_1(x) = \frac{x}{b-a}, \qquad y_2(x) = \frac{(b-h)(a-h)}{x(b-a)}$$

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. 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 $x_0 = \sqrt{(b-h)(a-h)}$.

We have determined the ideal distance of the tourist from the wall in two different ways. The angle corresponding to x_0 is

$$\varphi_0 = \arctan \frac{x_0(b-a)}{x_0^2 + (b-h)(a-h)} = \arctan \frac{b-a}{2\sqrt{(b-h)(a-h)}}.$$

When looking at the painting from the ground (being an ant, for instance), we have $h = 0$, and so

$$x_0 = \sqrt{ab}, \qquad \varphi_0 = \arctan \frac{b-a}{2\sqrt{ab}}.$$

If the painting is 1 meter high and its lower edge is 2 meters above ground ($a = 2, b = 3$), then the ant will see the painting at the largest angle $\varphi_0 \doteq 0.201\,4$ rad $\approx 11.5\,°$ at the distance $x_0 \doteq 2.45$ meters from the wall. If this 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, then the father should stand $x_0 \doteq 0.49$ meters from the wall and his son $x_0 \doteq 1.41$ meters, then. We can notice that the father has $\varphi_0 \doteq 0.795\,6$ rad $\approx 45.6\,°$ whereas his son has $\varphi_0 \doteq 0.339\,8$ rad $\approx 19.5\,°$. The quotient

$$\frac{0.795\,6}{0.339\,8} \approx \frac{45.6}{19.5} \doteq 2.3$$

proves what a strongly better view the father has. $\qquad\square$

5.F.10. Snell's law. Determine the refracted light ray between the point A in a homogeneous space with speed of light v_1 and the point B in a homogeneous space with speed of light v_2. See the picture.

Solution. Once again, we will omit units of measurement. We can assume that distances are given in meters, speeds v_1, v_2 in meters per second (and time in seconds, then). The ray is determined by Fermat's principle of least time: of 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 spaces, the ray will be a straight line (in this case, we will consider its segment). So it suffices to determine the

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:

$$\left(1 + \frac{x}{12}\right)^{12}, \quad \left(1 + \frac{x}{365}\right)^{365}, \quad \left(1 + \frac{x}{8760}\right)^{8760}, \cdots$$

Therefore, we could guess that

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n.$$

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:

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:

$$(1+b)^{k+1} = (1+b)^k(1+b) > (1+kb)(1+b)$$
$$= 1 + (k+1)b + kb^2 > 1 + (k+1)b$$

The statement is, of course, true for $b = -1$ as well. $\qquad\square$

point R (given by the value x) where the ray refracts. The distance between the points A and R is $\sqrt{h_1^2 + x^2}$, between points R and B it is $\sqrt{h_2^2 + (d-x)^2}$, then. The total time of the transmission of energy between the points A and B is thus given by the function

$$T(x) = \frac{\sqrt{h_1^2 + x^2}}{v_1} + \frac{\sqrt{h_2^2 + (d-x)^2}}{v_2}$$

in the variable $x \in [0, d]$. Let us emphasize that we want to find the point $x \in [0, d]$ at which the value $T(x)$ is minimal.

The derivative

$$T'(x) = \frac{x}{v_1 \sqrt{h_1^2 + x^2}} - \frac{d-x}{v_2 \sqrt{h_2^2 + (d-x)^2}}$$

is a continuous function on the interval $[0, d]$, so its sign can be easily described by its zero points. From the equation

$$T'(x) = 0, \quad \text{i. e.} \quad \frac{x}{v_1 \sqrt{h_1^2 + x^2}} = \frac{d-x}{v_2 \sqrt{h_2^2 + (d-x)^2}},$$

it follows that

$$\frac{\frac{x}{\sqrt{h_1^2 + x^2}}}{\frac{d-x}{\sqrt{h_2^2 + (d-x)^2}}} = \frac{v_1}{v_2}.$$

This expression is useful for us because (see the picture)

$$\sin \varphi_1 = \frac{x}{\sqrt{h_1^2 + x^2}}, \qquad \sin \varphi_2 = \frac{d-x}{\sqrt{h_2^2 + (d-x)^2}}.$$

Thus there is at most one stationary point; it is determined by

(1) $$\frac{\sin \varphi_1}{\sin \varphi_2} = \frac{v_1}{v_2}.$$

Let us realize that as $\varphi_1 \in [0, \pi/2]$ increases (when x increases), the angle $\varphi_2 \in [0, \pi/2]$ decreases. The sine is non-negative and increasing on the interval $[0, \pi/2]$, 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 there is the global minimum at the stationary point x_0.

Let us summarize the preceding: The ray is given by the point R of refraction (i. e. the value x_0), and the point R is given by the identity (1), which is called Snell's law in physics.

The quotient of v_1 and v_2 is constant for the given homogeneous spaces and determines an important quantity which describes the interface of optical spaces. It is called a refractive index and denoted by n. Usually, the first space is vacuum, i. e. $v_1 = c$, and $v_2 = v$, thus obtaining the (absolute) index of refraction $n = c/v$. For vacuum, we get $n = 1$, of course. This value is also used for air since its refractive index at the standard conditions (i. e. pressure of 101 325 Pa, temperature of 293 K and absolute humidity of $0.9 \, \mathrm{g \, m^{-3}}$) is

Now

$$\frac{a_n}{a_{n-1}} = \frac{(1 + \frac{1}{n})^n}{(1 + \frac{1}{n-1})^{n-1}} = \frac{(n^2 - 1)^n n}{n^{2n}(n-1)}$$

$$= \left(1 - \frac{1}{n^2}\right)^n \frac{n}{n-1} > \left(1 - \frac{1}{n}\right) \frac{n}{n-1} = 1.$$

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,

$$\frac{b_n}{b_{n+1}} = \frac{n}{n+1}\left(\frac{\frac{n+1}{n}}{\frac{n+2}{n+1}}\right)^{n+2} = \frac{n}{n+1}\left(\frac{n^2 + 2n + 1}{n^2 + 2n}\right)^{n+2}$$

$$= \frac{n}{n+1}\left(1 + \frac{1}{n(n+2)}\right)^{n+2}$$

$$\geq \frac{n}{n+1}\left(1 + \frac{n+2}{n(n+2)}\right) = 1.$$

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

$$\lim_{n \to \infty} b_n = \lim_{n \to \infty} (1 + \frac{1}{n}) a_n = \lim_{n \to \infty} a_n.$$

This limit determines one of the most important numbers in mathematics (besides the numbers 0, 1, and π), namely *Euler's number*[14] e. Thus

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n.$$

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

$$e^x = \lim_{n \to \infty} (a_n)^x.$$

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

$$e = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^{\frac{n}{x}}, \quad e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.$$

---

[14]The ingenious Swiss mathematician, physicist, astronomer, logician and engineer Leonhard Euler (1707-1783) was behind extremely many inventions, including original mathematical techniques and tools.

$n \doteq 1.000272$. Other spaces have $n > 1$ ($n = 1.31$ for ice, $n = 1.33$ for water, $n = 1.5$ 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 $n \doteq 1.331$ to $n \doteq 1.344$), where the index ordinarily decreases as the wave length increases. The speed of light in an optical space having $n > 1$ depends on its frequency. 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. We can further remind the well-known Newton's experiment with a glass prism from 1666.

Eventually, let us remark that our task always has a solution because we can choose the point R arbitrarily. If, together with the speeds v_1 and v_2, the angle φ_1 were given as well (our task could then be to calculate where the ray going from the point A intersects the line $y = c$ for a certain $c < 0$ when the interface of optical spaces is on the x-axis), then the angle $\varphi_2 \in (0, \pi/2)$ satisfying (1) might not exist. This corresponds to the total reflection (there is no refracted light at all). $\square$

Further miscellaneous problems concerning extrema of functions of a real variable can be found at **??**

5.F.11. Prove, that the polynomial $P(x) = x^5 - x^4 + 2x^3 - x^2 + x + 1$ has exactly one real root.

Solution. Any odd degree polynomial has at least one real root, since for big (in absolute value) negative x are the values $P(x)$ big (in absolute value) negative, for big positive x are the values $P(x)$ big positive and since $P(x)$ is a continuous function, it must attain the zero value. One can also argue with the fundamental theorem of algebra (12.2.8). We know, that the polynomial $P(x)$ has to have five roots over the field of complex numbers and that complex roots of the polynomial with real coefficients come in pairs of conjugated numbers. Therefore the polynomial has to have at least one real root.

If there were at least two roots, then according to the mean value theorem (it suffices the Rolle's version, cf. 5.3.8) there must be a $c \in (a, b)$, such that $P'(c) = 0$. But $P'(x) = 5x^4 - 4x^3 + 6x^2 - 2x + 1 = 2x^2(x-1)^2 + 3x^4 + 3x^2 + (x-1)^2 > 0$. Thus the polynomial has exactly one real root. $\square$

Denote the n-th term of this sequence by $u_n(x) = (1 + \frac{x}{n})^n$ and express it by the binomial theorem:

(1)
$$u_n(x) = 1 + n\frac{x}{n} + \frac{n(n-1)x^2}{2!n^2} + \cdots + \frac{n!x^n}{n!n^n}$$
$$= 1 + x + \frac{x^2}{2!}\left(1 - \frac{1}{n}\right)$$
$$+ \frac{x^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \ldots$$
$$+ \frac{x^n}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right).$$

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 $\frac{1}{k!}x^k$, $k = 0, 1, \ldots$. 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 $\lim_{k \to \infty} v_k(x)$ of the partial sums in the expression*

$$e^x = 1 + x + \frac{1}{2}x^2 + \cdots + \frac{1}{n!}x^n + \cdots = \sum_{i=0}^{\infty} \frac{1}{n!}x^n.$$

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. Fix x and recall that $v_n(x)$ is the sequences defined as the sums of the first n terms of the formal infinite expression

$$\sum_{j=0}^{\infty} c_j = \sum_{j=0}^{\infty} \frac{1}{j!}x^j.$$

While $v_n(x)$ is strictly increasing if $x > 0$, for all $m > n$

$$|v_m(x) - v_n(x)| \le \sum_{k=n+1}^{m} \frac{1}{k!}|x|^k = v_m(|x|) - v_n(|x|).$$

Hence $v_n(x)$ is a Cauchy sequence, and thus convergent, if $v_n(x)$ is always bounded (and thus convergent) for $x > 0$.

The quotient of adjacent terms in the series is $c_{j+1}/c_j = x/(n+1)$. Thus for every fixed x, there is a number $N \in \mathbb{N}$ such that $|c_{j+1}/c_j| < 1/2$ for all $j \ge N$. 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 $(1-q)(1 + q + \cdots + q^k) = 1 - q^{k+1}$. This means that the partial sums of the first $n > N$ terms of our formal sum with $x > 0$ can be estimated as follows

$$\left|v_n(x) - \sum_{j=0}^{N-1} \frac{1}{j!}x^j\right| < \frac{1}{N!}x^N \sum_{j=0}^{n-N} \frac{1}{2^j}.$$

5.F.12. Let $f : \mathbb{R} \to (0, \infty)$ be a continuously differentiable function. Prove that there exists $\xi \in (0, 1)$ such that

$$e^{f'(\xi)} f(0)^{f(xi)} = f(1)^{f(\xi)}.$$

Solution. We can equivalently transform the equation:

$$e^{f'(\xi)} = \left(\frac{f(1)}{f(0)} \right)^{f(\xi)}$$

$$\frac{f'(\xi)}{f'(\xi)} = \ln f(1) - \ln f(0).$$

The existence of such a x is guaranteed by the Lagrage's mean value theorem (cf. 5.3.9) for the function $g(x) = \ln(f(x))$ (there is $g'(x) = \frac{f'(x)}{f(x)}$). $\square$

G. L'Hospital's rule

5.G.1. Verify that the limit

(a)
$$\lim_{x \to 0} \frac{\sin(2x) - 2\sin x}{2e^x - x^2 - 2x - 2} \quad \text{is of the type } \frac{0}{0};$$

(b)
$$\lim_{x \to 0+} \frac{\ln x}{\cot x} \quad \text{is of the type } \frac{\infty}{\infty};$$

(c)
$$\lim_{x \to 1+} \left(\frac{x}{x-1} - \frac{1}{\ln x} \right) \quad \text{is of the type } \infty - \infty;$$

(d)
$$\lim_{x \to 1+} \left(\ln(x-1) \cdot \ln x \right) \quad \text{is of the type } 0 \cdot \infty;$$

(e)
$$\lim_{x \to 0+} (\cot x)^{\frac{1}{\ln x}} \quad \text{is of the type } \infty^0;$$

(f)
$$\lim_{x \to 0} \left(\frac{\sin x}{x} \right)^{\frac{1}{x^2}} \quad \text{is of the type } 1^\infty;$$

(g)
$$\lim_{x \to 1-} \left(\cos \frac{\pi x}{2} \right)^{\ln x} \quad \text{is of the type } 0^0.$$

Then calculate it using l'Hospital's rule.
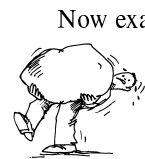
Solution. We can immediately assert that

(a)
$$\lim_{x \to 0} (\sin(2x) - 2\sin x) = 0 - 0 = 0,$$
$$\lim_{x \to 0} (2e^x - x^2 - 2x - 2) = 2 - 0 - 0 - 2 = 0;$$

(b)
$$\lim_{x \to 0+} \ln x = -\infty, \qquad \lim_{x \to 0+} \cot x = +\infty;$$

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.

Now examine the sequence of numbers u_n, whose limit is e^x. Consider $n > N$ for some fixed N (imagine already N is very large) and choose a fixed number $k < N$. Write $u_{n,k}$ for the sum of the first k summands in the expression (1) for u_n. Having fixed x and some $\varepsilon > 0$, we can choose k big enough to ensure $|u_{n,k} - u_n| < \varepsilon$ (remind u_n is a convergent and thus Cauchy sequence). Indeed, the absolute values of the omitted terms are less than those in v_n.

If N is large enough, $|u_{n,k} - v_k| < \varepsilon$ for all $n > N$. Indeed, there is only a fixed number of the brackets in the summands of $u_{n,k}$ and they will all be arbitrarily close to 1, if n is large enough.

Summarizing, for each $\varepsilon > 0$ the choices of k and n lead to the estimate $|v_k - u_n| \leq |v_k - u_{n,k}| + |u_{n,k} - u_n| < 2\varepsilon$. Choosing a sequence of $\varepsilon_i = 1/2i$, we find subsequences v_{k_i} and u_{n_i} satisfying $|v_{k_i} - u_{n_i}| < \frac{1}{i}$. Thus the two convergent sequences must both have the same limit:

$$\lim_{k \to \infty} v_k = \lim_{n \to \infty} u_n.$$

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 remains to compute

$$\lim_{x \to 0} \frac{(1 + x + \frac{1}{2}x^2 + \dots) - 1}{x} = \lim_{x \to 0} \frac{x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \dots}{x}.$$

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):

$$\lim_{x \to 0} \left(\lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{k!} x^{k-1} \right) = 1 + \lim_{x \to 0} \left(\lim_{n \to \infty} \sum_{k=2}^{n} \frac{1}{k!} x^{k-1} \right).$$

Next, for each $\varepsilon > 0$ we can find N such that $\lim_{n \to \infty} |\sum_{k=N}^{n} \frac{1}{k!} x^k| \leq \lim_{n \to \infty} \sum_{k=N}^{n} \frac{1}{k!} < \varepsilon$ for all $x \in [-1, 1]$. Next we restrict the interval for x enough to ensure that the remaining first terms yield $\sum_{k=2}^{N-1} \frac{1}{k!} x^k < \varepsilon$, too. 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.

(c)

$$\lim_{x \to 1+} \frac{x}{x-1} = +\infty, \qquad \lim_{x \to 1+} \frac{1}{\ln x} = +\infty;$$

(d)

$$\lim_{x \to 1+} \ln x = 0, \qquad \lim_{x \to 1+} \ln(x-1) = -\infty;$$

(e)

$$\lim_{x \to 0+} \cot x = +\infty, \qquad \lim_{x \to 0+} \frac{1}{\ln x} = 0;$$

(f)

$$\lim_{x \to 0} \frac{\sin x}{x} = 1, \qquad \lim_{x \to 0} \frac{1}{x^2} = +\infty;$$

(g)

$$\lim_{x \to 1-} \cos \frac{\pi x}{2} = 0, \qquad \lim_{x \to 1-} \ln x = 0.$$

The case (a). Applying l'Hospital's rule transforms the limit

$$\lim_{x \to 0} \frac{\sin(2x) - 2\sin x}{2e^x - x^2 - 2x - 2}$$

into the limit

$$\lim_{x \to 0} \frac{2\cos(2x) - 2\cos x}{2e^x - 2x - 2},$$

which is of the type $0/0$. Two more applications of the rule lead to

$$\lim_{x \to 0} \frac{-4\sin(2x) + 2\sin x}{2e^x - 2}$$

and (the above limit is also of the type $0/0$)

$$\lim_{x \to 0} \frac{-8\cos(2x) + 2\cos x}{2e^x} = \frac{-8+2}{2} = -3.$$

Altogether, we have (returning to the original limit)

$$\lim_{x \to 0} \frac{\sin(2x) - 2\sin x}{2e^x - x^2 - 2x - 2} = -3.$$

Let us remark that multiple application of l'Hospital's rule in an exercise is quite common.

From now on, we will set the limits of quotients of derivatives obtained by l'Hospital's rule equal the limits of the original quotients. We can do this if the gained limits on the right sides really exist, i. e. actually we will make sure that what we write is senseful only afterwards.

The case (b). This time, differentiation of the numerator and the denominator gives

$$\lim_{x \to 0+} \frac{\ln x}{\cot x} = \lim_{x \to 0+} \frac{\frac{1}{x}}{\frac{-1}{\sin^2 x}} = \lim_{x \to 0+} \frac{-\sin^2 x}{x}.$$

The last limit can be determined easily (we even know it). From

$$\lim_{x \to 0+} -\sin x = 0, \qquad \lim_{x \to 0+} \frac{\sin x}{x} = 1,$$

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

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots + a_k + \ldots,$$

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

$$s = \lim_{k \to \infty} s_n$$

of the partial sums exists and is finite.

If the sequence of partial sums has an improper limit, the series *diverges* to ∞ or $-\infty$. If the limit of the partial sums does not exist, the series *oscillates*.

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

$$|s_m - s_n| = |a_{n+1} + \cdots + a_m|$$

must be arbitrarily small for sufficiently large $m > n$. Since

$$|a_{n+1}| + \cdots + |a_m| > |a_{n+1} + \cdots + a_m|,$$

the convergence of the series $\sum_{k=0}^{\infty} |a_n|$ implies the convergence of the series $\sum_{k=0}^{\infty} a_n$.

ABSOLUTELY CONVERGENT SERIES

A series $\sum_{k=0}^{\infty} a_n$ is *absolutely convergent* if the series $\sum_{n=0}^{\infty} |a_n|$ 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.

the result $0 = 0 \cdot 1$ follows. We could also have used l'Hospital's rule again (now for the expression $0/0$), obtaining the result

$$\lim_{x\to 0+} \frac{-\sin^2 x}{x} = \lim_{x\to 0+} \frac{-2\cdot\sin x\cdot\cos x}{1} = \frac{-2\cdot 0\cdot 1}{1} = 0.$$

The case (c). By mere transforming to a common denominator:

$$\lim_{x\to 1+}\left(\frac{x}{x-1}-\frac{1}{\ln x}\right) = \lim_{x\to 1+}\frac{x\ln x - (x-1)}{(x-1)\ln x}$$

we have obtained the type $0/0$. We have that

$$\lim_{x\to 1+}\frac{x\ln x-(x-1)}{(x-1)\ln x} = \lim_{x\to 1+}\frac{\ln x + \frac{x}{x}-1}{\frac{x-1}{x}+\ln x}$$
$$= \lim_{x\to 1+}\frac{\ln x}{1-\frac{1}{x}+\ln x}.$$

We have the quotient $0/0$, which (again by l'Hospital's rule) satisfies

$$\lim_{x\to 1+}\frac{\ln x}{1-\frac{1}{x}+\ln x} = \lim_{x\to 1+}\frac{\frac{1}{x}}{\frac{1}{x^2}+\frac{1}{x}} = \frac{1}{1+1} = \frac{1}{2}.$$

Returning to the original limit, we write the result

$$\lim_{x\to 1+}\left(\frac{x}{x-1}-\frac{1}{\ln x}\right) = \frac{1}{2}.$$

The case (d). We transform the assigned expression into the type ∞/∞ (to be precise, into the type $-\infty/\infty$) by creating the fraction

$$\lim_{x\to 1+}\ln(x-1)\cdot\ln x = \lim_{x\to 1+}\frac{\ln(x-1)}{\frac{1}{\ln x}}.$$

By l'Hospital's rule,

$$\lim_{x\to 1+}\frac{\ln(x-1)}{\frac{1}{\ln x}} = \lim_{x\to 1+}\frac{\frac{1}{x-1}}{-\frac{1}{\ln^2 x}\cdot\frac{1}{x}} = \lim_{x\to 1+}\frac{-x\ln^2 x}{x-1}.$$

This indeterminate form (of the type $0/0$) can once again be determined by l'Hospital's rule:

$$\lim_{x\to 1+}\frac{-x\ln^2 x}{x-1} = \lim_{x\to 1+}\frac{-\ln^2 x - 2x\ln x\cdot\frac{1}{x}}{1} = -\frac{0+0}{1} = 0.$$

The cases (e), (f), (g). Since

$$\lim_{x\to 0+}(\cot x)^{\frac{1}{\ln x}} = e^{\lim_{x\to 0+}\frac{\ln(\cot x)}{\ln x}};$$

$$\lim_{x\to 0}\left(\frac{\sin x}{x}\right)^{\frac{1}{x^2}} = e^{\lim_{x\to 0}\frac{\ln\frac{\sin x}{x}}{x^2}};$$

$$\lim_{x\to 1-}\left(\cos\frac{\pi x}{2}\right)^{\ln x} = e^{\lim_{x\to 1-}\left(\ln x\cdot\ln\left(\cos\frac{\pi x}{2}\right)\right)},$$

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 $\sigma : \mathbb{N}\to\mathbb{N}$ of integers.

PROOF. Both the first and the second statements are a straightforward consequence of the corresponding properties of limits. 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,

$$\left(\sum_{n=0}^{k}a_n\right)\cdot\left(\sum_{n=0}^{k}b_n\right) \to \left(\sum_{n=0}^{\infty}a_n\right)\cdot\left(\sum_{n=0}^{\infty}b_n\right).$$

Thus it suffices to prove that

$$0 = \lim_{k\to\infty}\left(\left(\sum_{n=0}^{k}a_n\right)\cdot\left(\sum_{n=0}^{k}b_n\right) - \sum_{n=0}^{k}c_n\right).$$

Consider the expressions

$$\left(\sum_{n=0}^{k}a_n\right)\cdot\left(\sum_{n=0}^{k}b_n\right) = \sum_{0\le i,j\le k}a_i b_j,$$

$$c_n = \sum_{\substack{i+j=n\\0\le i,j\le k}}a_i b_j, \quad \sum_{n=0}^{k}c_n = \sum_{\substack{i+j\le k\\0\le i,j\le k}}a_i b_j.$$

along with the bound

$$\left|\left(\sum_{n=0}^{k}a_n\right)\cdot\left(\sum_{n=0}^{k}b_n\right) - \sum_{n=0}^{k}c_n\right| = \left|\sum_{\substack{i+j>k\\0\le i,j\le k}}a_i b_j\right|$$

$$\le \sum_{\substack{i+j>k\\0\le i,j\le k}}|a_i b_j|.$$

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

it suffices to calculate the limits given in the argument of the exponential function. By l'Hospital's rule and simple rearrangements, we get

$$\lim_{x\to 0+} \frac{\ln(\cot x)}{\ln x} \left[\text{type } \frac{+\infty}{-\infty}\right] = \lim_{x\to 0+} \frac{\frac{1}{\cot x}\cdot\frac{-1}{\sin^2 x}}{\frac{1}{x}}$$

$$= \lim_{x\to 0+} \frac{-x}{\cos x\cdot\sin x}\left[\text{type }\frac{0}{0}\right] = \lim_{x\to 0+}\frac{-1}{\cos^2 x - \sin^2 x}$$

$$= \frac{-1}{1-0} = -1;$$

$$\lim_{x\to 0}\frac{\ln\frac{\sin x}{x}}{x^2}\left[\text{type }\frac{0}{0}\right] = \lim_{x\to 0}\frac{\frac{x}{\sin x}\cdot\frac{x\cos x - \sin x}{x^2}}{2x}$$

$$= \lim_{x\to 0}\frac{x\cos x - \sin x}{2x^2\sin x}\left[\text{type }\frac{0}{0}\right]$$

$$= \lim_{x\to 0}\frac{\cos x - x\sin x - \cos x}{4x\sin x + 2x^2\cos x}$$

$$= \lim_{x\to 0}\frac{-\sin x}{4\sin x + 2x\cos x}\left[\text{type }\frac{0}{0}\right]$$

$$= \lim_{x\to 0}\frac{-\cos x}{4\cos x + 2\cos x - 2x\sin x} = \frac{-1}{4+2-0} = -\frac{1}{6},$$

hence

$$\lim_{x\to 0+}(\cot x)^{\frac{1}{\ln x}} = e^{-1} = \frac{1}{e};$$

$$\lim_{x\to 0}\left(\frac{\sin x}{x}\right)^{\frac{1}{x^2}} = e^{-\frac{1}{6}} = \frac{1}{\sqrt[6]{e}}.$$

We can proceed similarly when determining the last limit. We have that

$$\lim_{x\to 1-}(\ln x)\cdot\ln\left(\cos\frac{\pi x}{2}\right)$$

$$= \lim_{x\to 1-}\frac{\ln\left(\cos\frac{\pi x}{2}\right)}{\frac{1}{\ln x}} = \left[\text{type }\frac{-\infty}{-\infty} = \frac{\infty}{\infty}\right]$$

$$= \lim_{x\to 1-}\frac{\frac{1}{\cos\frac{\pi x}{2}}\left(-\sin\frac{\pi x}{2}\right)\frac{\pi}{2}}{-\frac{1}{\ln^2 x}\cdot\frac{1}{x}}$$

$$= \frac{\pi}{2}\lim_{x\to 1-}\frac{x\sin\frac{\pi x}{2}\cdot\ln^2 x}{\cos\frac{\pi x}{2}}.$$

Since this form is of the type $0/0$, we could continue by using l'Hospital's rule; instead, we will go from

$$\lim_{x\to 1-}\frac{x\sin\frac{\pi x}{2}\cdot\ln^2 x}{\cos\frac{\pi x}{2}}$$

over to the product of limits

$$\lim_{x\to 1-}\left(x\sin\frac{\pi x}{2}\right)\cdot\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}} = 1\cdot\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}}.$$

Only now we apply l'Hospital's rule for

$$\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}} = \left[\text{type }\frac{0}{0}\right] = \lim_{x\to 1-}\frac{2\ln x\cdot\frac{1}{x}}{\left(-\frac{\pi}{2}\right)\sin\frac{\pi x}{2}} = \frac{0}{-\frac{\pi}{2}} = 0.$$

the product and remove only those whose indices are both at most $k/2$.

$$\sum_{\substack{i+j>k\\0\le i,j\le k}}|a_i b_j| \le \sum_{0\le i,j\le k}|a_i b_j| - \sum_{0\le i,j\le k/2}|a_i b_j|.$$

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 α such that for all $N > \alpha$ both estimates are true:

$$\sum_{n=N}^{\infty}|a_n| < \varepsilon, \qquad \left|\sum_{n=0}^{N}a_n - S\right| < \varepsilon.$$

Now, consider any permutation σ of the indices and write $I_\sigma = \{\sigma^{-1}(0),\ldots,\sigma^{-1}(\alpha)\}$. Then, for each $N > \max I_\sigma$ clearly

$$\left|\sum_{n=0}^{N}a_{\sigma(n)} - S\right| = \left|\sum_{n\in I_\sigma}^{N}a_{\sigma(n)} - S + \sum_{n\notin I_\sigma}^{N}a_{\sigma(n)}\right|$$

$$\le \left|\sum_{n=0}^{\alpha}a_n - S\right| + \sum_{n\notin I_\sigma}^{N}|a_{\sigma(n)}|$$

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 2ε. 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.

Altogether, we have

$$\lim_{x\to 1-} \left(\ln x \cdot \ln \left(\cos \frac{\pi x}{2} \right) \right) = \frac{\pi}{2} \cdot 1 \cdot 0 = 0,$$

i. e.

$$\lim_{x\to 1-} \left(\cos \frac{\pi x}{2} \right)^{\ln x} = e^0 = 1.$$

$\square$

5.G.2. As we have implicitly mentioned, using l'Hospital's rule can lead to a non-existing limit even though the original limit exists: Determine the limit

$$\lim_{x\to\infty} \frac{x + \sin x}{x}$$

Solution. The limit is of the type $\frac{\infty}{\infty}$, by l'Hospital's rule, we get that

$$\lim_{x\to\infty} \frac{x + \sin x}{x} = \lim_{x\to\infty} \frac{1 + \cos x}{1},$$

and since the limit $\lim_{x\to\infty} \cos x$ does not exist, nor does the limit $\lim_{x\to\infty} 1 + \cos x$. However, the original limit exists because

$$\frac{x - 1}{x} \leq \frac{x + \sin x}{x} \leq \frac{x + 1}{x},$$

and by the squeeze theorem,

$$1 = \lim_{x\to\infty} \frac{x + \sin x}{x} \leq \lim_{x\to\infty} \frac{x + \sin x}{x} \leq \lim_{x\to\infty} \frac{x + 1}{x} = 1.$$

$\square$

5.G.3. Determine

$$\lim_{x\to+\infty} \frac{\ln x}{x}, \quad \lim_{x\to 0+} x \ln \frac{1}{x}, \quad \lim_{x\to 0+} x \, e^{\frac{1}{x}},$$

$$\lim_{x\to 0-} x \, e^{-\frac{1}{x}}, \quad \lim_{x\to 0} \frac{e^{-\frac{1}{x^2}}}{x^{100}}, \quad \lim_{x\to+\infty} \left(\ln x - x \right),$$

$$\lim_{x\to+\infty} \frac{x}{x + \ln x \cdot \cos x}, \quad \lim_{x\to+\infty} \frac{\sqrt[3]{x + 1}}{\sqrt[5]{x + 3}}, \quad \lim_{x\to+\infty} \frac{x}{\sqrt{x^2 + 1}}.$$

Solution. It can easily be shown (for instance, by n-fold use of l'Hospital's rule) that for any $n \in \mathbb{N}$, it holds that

$$\lim_{x\to+\infty} \frac{x^n}{e^x} = 0, \quad \text{i. e.} \quad \lim_{x\to+\infty} \frac{e^x}{x^n} = +\infty.$$

The squeeze theorem implies the following generalization for real numbers $a > 0$:

$$\lim_{x\to+\infty} \frac{x^a}{e^x} = 0, \quad \text{i. e.} \quad \lim_{x\to+\infty} \frac{e^x}{x^a} = +\infty.$$

Taking into account that the graphs of the functions $y = e^x$ and $y = \ln x$ (the inverse function to $y = e^x$) are symmetric with regard to the line $y = x$, we further see that

$$\lim_{x\to+\infty} \frac{\ln x}{x} = 0, \quad \text{i. e.} \quad \lim_{x\to+\infty} \frac{x}{\ln x} = +\infty.$$

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 $\lim_{n\to\infty} a_n = 0$.

(2) (The comparison test) If T converges and $|a_n| < b_n$, then S converges absolutely. If $b_n < |a_n|$ 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 $\lim_{n\to\infty} \frac{a_n}{b_n} = r > 0$ 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

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = q.$$

Then the series S converges absolutely for $|q| < 1$ and does not converge for $|q| > 1$. If $|q| = 1$ the series may or may not converge.

(5) (The root test) If the limit

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = q$$

exists, then the series converges absolutely for $q < 1$. It does not converge 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 $\lim_{n\to\infty} a_n$ 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 ε. Thus the sequence of partial sums cannot be a Cauchy sequence and, therefore, it cannot be convergent.

(2) The result is a straightforward consequence of the squeeze theorem, cf. 5.2.12.

(3) Since limit $r = \lim_{n\to\infty} \frac{a_n}{b_n}$ 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,

$$a_{j+1} < r \cdot a_j \leq r^{(j-N+1)} a_N,$$

Thus we have obtained the first result. That could also be derived from l'Hospital's rule because

$$\lim_{x \to +\infty} \frac{\ln x}{x} = \lim_{x \to +\infty} \frac{\frac{1}{x}}{1} = \lim_{x \to +\infty} \frac{1}{x} = 0.$$

Let us point out that l'Hospital's rule can be used to calculate all of the following five limits. However, it is possible to determine these limits by much simpler means. For instance, the substitution $y = 1/x$ leads to

$$\lim_{x \to 0+} x \ln \frac{1}{x} = \lim_{y \to +\infty} \frac{\ln y}{y} = 0;$$

$$\lim_{x \to 0+} x\, e^{\frac{1}{x}} = \lim_{y \to +\infty} \frac{e^y}{y} = +\infty.$$

Of course, $x \to 0+$ gives $y = 1/x \to +\infty$ (we write $1/+0 = +\infty$).

By the substitutions $u = -1/x$, $v = 1/x^2$ we get that, respectively,

$$\lim_{x \to 0-} x\, e^{-\frac{1}{x}} = \lim_{u \to +\infty} -\frac{e^u}{u} = -\infty;$$

$$\lim_{x \to 0} \frac{e^{-\frac{1}{x^2}}}{x^{100}} = \lim_{v \to +\infty} \frac{v^{50}}{e^v} = 0,$$

where $x \to 0-$ corresponds to $u = -1/x \to +\infty$ (we write $-1/-0 = +\infty$) and then $x \to 0$ to $v = 1/x^2 \to +\infty$ (again $1/+0 = +\infty$). We have also clarified that

$$\lim_{x \to +\infty} (\ln x - x) = \lim_{x \to +\infty} -x = -\infty.$$

Potential doubts can be scattered by the limit

$$\lim_{x \to +\infty} \frac{\ln x - x}{\ln x} = \lim_{x \to +\infty} \left(1 - \frac{x}{\ln x}\right) = -\infty,$$

which proves that even when decreasing the absolute value of the considered expression (without changing the sign), the absolute value of the expression remains unbounded.

We can equally easily determine

$$\lim_{x \to +\infty} \frac{x}{x + \ln x \cdot \cos x} = \lim_{x \to +\infty} \frac{x}{x} = 1;$$

$$\lim_{x \to +\infty} \frac{\sqrt[3]{x+1}}{\sqrt[5]{x+3}} = \lim_{x \to +\infty} \frac{\sqrt[3]{x}}{\sqrt[5]{x}} = +\infty;$$

$$\lim_{x \to +\infty} \frac{x}{\sqrt{x^2+1}} = \lim_{x \to +\infty} \frac{x}{\sqrt{x^2}} = 1.$$

We have seen that the l'Hospital's rule may not be the best method for calculating limits of types $0/0$, ∞/∞. The three preceding exercises illustrate that it even cannot be applied in all cases (for indeterminate forms). If we had applied it to the first problem, we would have obtained, for $x > 0$, the quotient

$$\frac{1}{1 + \frac{\cos x}{x} - \ln x \cdot \sin x} = \frac{x}{x + \cos x - x \ln x \cdot \sin x},$$

where the last equality follows from the general equality $(1 - r)(1 + r^2 + \cdots + r^k) = 1 - r^{k+1}$. But this means that the partial sums s_n are, for large $n > N$, bounded from above by the sums

$$s_n < \sum_{j=0}^{N} a_j + a_N \sum_{j=0}^{n-N} r^j = \sum_{j=0}^{N} a_j + a_N \frac{1 - r^{n-N+1}}{1 - r}.$$

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. □

In the proofs of the last two statements of the 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 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 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 called *upper limit* of the sequence and denoted by

$$\limsup_{n \to \infty} b_n.$$

The advantage is that the upper limit always exists. 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

$$q = \limsup_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right|,$$

then the series S converges absolutely for $q < 1$ and does not converge for $q > 1$. For $q = 1$, it may or may not converge.

(2) If

$$q = \limsup_{n \to \infty} \sqrt[n]{|a_n|},$$

the series converges absolutely for $q < 1$ while it does not converge for $q > 1$. For $q = 1$, it may or may not converge.

which is more complicated than the original one. The limit for $x \to +\infty$ does not even exist, so one of the prerequisites of l'Hospital's rule is not satisfied. In the second case, any number of multiple uses of l'Hospital's rule leads to indeterminate forms. For the last problem, l'Hospital's rule sends us back to the original limit: first it gives the fraction

$$\frac{1}{\frac{2x}{2\sqrt{x^2+1}}} = \frac{\sqrt{x^2+1}}{x}$$

and then

$$\frac{\frac{2x}{2\sqrt{x^2+1}}}{1} = \frac{x}{\sqrt{x^2+1}}.$$

From here, we can deduce that the limit equals 1 (we are looking for a non-negative real number $a \in \mathbb{R}$ such that $a = a^{-1}$) only if we have already shown it exists at all. □

Other examples concerning calculation of limits by l'Hospital's rule can be found at page .

H. Infinite series

Infinite series naturally appear in a series (of problems).

5.H.1. Sierpiński carpet. The unit squares is divided into

nine equal squares and the middle one is removed. Each of the eight remaining squares is again divided into nine equal subsquares and the middle subsquare (of each of the eight squares) is removed again. Having applied this procedure ad infinitum, determine the area of the resulting figure.
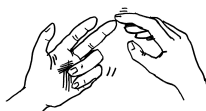
Solution. In the first step, a square having the area of $1/9$ is removed. In the second steps, eight squares (each having the area of 9^{-2}, i. e. totaling to $8 \cdot 9^{-2}$) are removed. Every further iteration removes eight times more squares than in the previous steps, but the squares are nine times smaller. The sum of areas of all the removed squares is

$$\frac{1}{9} + \frac{8}{9^2} + \frac{8^2}{9^3} + \cdots = \sum_{n=0}^{\infty} \frac{8^n}{9^{n+1}}.$$

The area of the remaining figure (known as Sierpiński carpet) thus equals

$$1 - \sum_{n=0}^{\infty} \frac{8^n}{9^{n+1}} = 1 - \frac{1}{9} \sum_{n=0}^{\infty} \left(\frac{8}{9}\right)^n = 1 - \frac{1}{9} \cdot \frac{1}{1-\frac{8}{9}} = 0.$$

□

5.H.2. 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, build another equilateral triangle above (i. e. pointing out

5.4.6. Alternating series. The condition $a_n \to 0$ is a necessary but not sufficient condition for the convergence of the series $\sum_{n=0}^{\infty} a_n$. However, there is the *Leibniz criterion* of convergence.

LEIBNIZ 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* $\lim_{n\to\infty} a_n = 0$. *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

$$s_{2(k+1)+1} = s_{2k+1} + a_{2k+2} - a_{2k+3} \geq s_{2k+1}$$
$$s_{2(k+1)} = s_{2k} - a_{2k+1} + a_{2k+2} \leq s_{2k}$$
$$s_{2k+1} - s_{2k} = -a_{2k+1} \to 0$$
$$s_2 \geq s_{2k} \geq s_{2k+1} \geq s_1.$$

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 to its supremum, while the even ones converge to the infimum. The previous line says they coincide, if and only if $\lim_{n\to\infty} a_n = 0$ which proves the first claim.

At the same the limit value a of the series is always at most s_{2k+1} and at least s_{2k}. Thus, the latter partial sums cannot differ by more than a_{2k+1}. □

Remark. As obvious from the latter theorem, convergent alternating series are often not convernging 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 series theorem saying that conditionaly convergent series can be braught to any finite or infite value by appropriate rearrangement of the terms in the sum. We shall not go into the proof here.

5.4.7. 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 $q = \limsup_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|$ or $q = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$, and $0 < q < 1$. In the estimate of the error of approximation of the limit s_∞ 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 non dependent of n. Thus the convergence rate is quite fast, in particular if

from, not into, the original triangle) the middle part and re-move the middle part. This transforms the original equilateral triangle into a six-pointed star. Once again, repeat this step ad infinitum, thus obtaining 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. Hence it follows that the snowflake's perimeter can be expressed as the limit

$$d_n = 3 \left(\tfrac{4}{3} \right)^n \quad \text{and} \quad \lim_{n \to \infty} d_n = +\infty.$$

The figure's area is apparently increasing during the construction. To determine it, it thus suffices to catch the rise between two consecutive steps. The number of the figure's sides is four times higher every step (the line segments are divided into thirds and one of them is doubled) and the new sides are three times shorter. The figure's area thus 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

$$S_0 + 3 \cdot \tfrac{S_0}{9}.$$

for the area of the six-pointed star. Similarly, in the next step we obtain the area of the figure as

$$S_0 + 3 \cdot \tfrac{S_0}{9} + 4 \cdot 3 \cdot \tfrac{S_0}{9^2}.$$

Now it is easy to deduce that the area of the resulting snowflake equals the limit

$$\lim_{n \to \infty} \left(S_0 + 3 \cdot \tfrac{S_0}{9} + 4 \cdot 3 \cdot \tfrac{S_0}{9^2} + \cdots + 4^n \cdot 3 \cdot \tfrac{S_0}{9^{n+1}} \right) =$$
$$S_0 \lim_{n \to \infty} \left(1 + \tfrac{1}{3} + \tfrac{1}{3} \cdot \tfrac{4}{9} + \cdots + \tfrac{1}{3} \cdot \left(\tfrac{4}{9} \right)^n \right) =$$
$$S_0 \left[1 + \tfrac{1}{3} \lim_{n \to \infty} \left(1 + \tfrac{4}{9} + \cdots + \left(\tfrac{4}{9} \right)^n \right) \right] =$$
$$S_0 \left[1 + \tfrac{1}{3} \lim_{n \to \infty} \sum_{k=0}^{n} \left(\tfrac{4}{9} \right)^k \right] = S_0 \left[1 + \tfrac{1}{3} \sum_{k=0}^{\infty} \left(\tfrac{4}{9} \right)^k \right] =$$
$$S_0 \left[1 + \tfrac{1}{3} \cdot \tfrac{1}{1 - \tfrac{4}{9}} \right] = \tfrac{8}{5} S_0.$$

The snowflake's area is thus equal to 8/5 of the area of the original triangle, i. e.

$$\tfrac{8}{5} S_0 = \tfrac{8}{5} \cdot \tfrac{\sqrt{3}}{4} = \tfrac{2\sqrt{3}}{5}.$$

k is much smaller than 1 (and we can get k 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.

5.4.8. Power series. If we consider not 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*

$$S(x) = \sum_{n=0}^{\infty} f_n(x).$$

POWER SERIES

A *power series* is a series of functions given by the expression

$$S(x) = \sum_{n=0}^{\infty} a_n x^n$$

with coefficients $a_n \in \mathbb{C}$, $n = 0, 1, \ldots$.

$S(x)$ 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 $|x| > \rho$.

5.4.9. 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. Notice that the upper limit $r = \limsup_{n \to \infty} \sqrt[n]{|a_n|}$ equals the limit $\lim_{n \to \infty} \sqrt[n]{|a_n|}$, whenever this limit exists.

Actually, we should also notice that our argument on the converengence 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 in Chapter 9.

CONVERGENCE AND DIFFERENTIATION

Theorem. *Let $S(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series and let*

$$r = \limsup_{n \to \infty} \sqrt[n]{|a_n|}.$$

Then the radius of convergence of the series S is $\rho = r^{-1}$.

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.

Let us notice that this snowflake is an example of an infinitely long curve which encloses a finite area. □

5.H.3. Show that the so-called *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

diverges.

Solution. For any natural number k, the sum of the first 2^k terms of this series is greater than $k/2$:

$$1 + \underbrace{\frac{1}{2}}_{>\frac{1}{2}} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{4}+\frac{1}{4}=\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}} + \dots$$

as the sum of the terms from $2^l + 1$ to 2^{l+1} is always greater than 2^l-times (its number) $1/2^l$ (the least one of them), which sums to $1/2$. □

5.H.4. Determine whether the following series converge, or diverge:

i) $\sum_{n=1}^{\infty} \frac{2^n}{n}$

ii) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$

iii) $\sum_{n=1}^{\infty} \frac{1}{n \cdot 2^{100000}}$

iv) $\sum_{n=1}^{\infty} \frac{1}{(1+i)^n}$

Solution.

i) We will examine the convergence by the ratio test:

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n\to\infty} \left| \frac{\frac{2^{n+1}}{n+1}}{\frac{2^n}{n}} \right| = \lim_{n\to\infty} \frac{2(n+1)}{n} = 2 > 1,$$

so the series diverges.

ii) We will bound the series from below: we know that $\frac{1}{n} \leq \frac{1}{\sqrt{n}}$ for any natural number n. Thus the sequence of the partial sums s_n of the examined series and the sequence of the partial sums s'_n of the harmonic series satisfy:

$$s_n = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \geq \sum_{i=1}^{n} \frac{1}{n} = s'_n.$$

Since the harmonic series diverges (see the previous exercise), by definition, the sequence of its partial sums $\{s'_n\}_{n=1}^{\infty}$ diverges as well. Therefore the sequence of its partial sums $\{s_n\}_{n=1}^{\infty}$ also diverges and so does the examined sequence.

iii) This series is divergent since it is a multiple of the harmonic series.

Calculate (if the limit exists)

$$\lim_{n\to\infty} \sqrt[n]{|a_n x^n|} = r|x|.$$

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 statements about continuity and the derivatives are proved later in a more general context, see . □

5.4.10. 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 is 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 1 because

$$\lim_{n\to\infty} \left| \frac{\frac{1}{n+1} x^{n+1}}{\frac{1}{n} x^n} \right| = |x| \lim_{n\to\infty} \left| \frac{n}{n+1} \right| = |x|.$$

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.

5.4.11. 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.

iv) The examined series is geometric, with common ratio $\frac{1}{1+i}$. Such a sequence is convergent if and only if the absolute value of the common ratio is less than one. We know that

$$\left|\frac{1}{1+i}\right| = \left|\frac{1-i}{2}\right| = \left|\frac{1}{2} - \frac{1}{2}i\right| = \sqrt{\frac{1}{4} + \frac{1}{4}} = \frac{\sqrt{2}}{2} < 1,$$

hence the series converges, and we are even able to calculate it:

$$\sum_{n=1}^{\infty} \frac{1}{(1+i)^n} = \frac{1}{1 - \frac{1}{1+i}} = \frac{1+i}{i} = 1 - i.$$

$\square$

5.H.5. Calculate the series

(a) $\sum_{n=1}^{\infty} \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}\right)$;

(b) $\sum_{n=0}^{\infty} \frac{5}{3^n}$;

(c) $\sum_{n=1}^{\infty} \left(\frac{3}{4^{2n-1}} + \frac{2}{4^{2n}}\right)$;

(d) $\sum_{n=1}^{\infty} \frac{n}{3^n}$;

(e) $\sum_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)}$.

Solution. The case (a). From the definition, the series is equal to

$$\sum_{n=1}^{\infty} \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}\right) =$$
$$\lim_{n\to\infty} \left(\left(\frac{1}{\sqrt{1}} - \frac{1}{\sqrt{2}}\right) + \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}}\right) + \cdots + \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}\right)\right)$$
$$= \lim_{n\to\infty} \left(1 + \left(-\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right) + \cdots + \left(-\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - \frac{1}{\sqrt{n+1}}\right)$$
$$= 1.$$

The case (b). Apparently, this sequence is a quintuple of the standard geometric series with the common ratio $q = 1/3$, hence

$$\sum_{n=0}^{\infty} \frac{5}{3^n} = 5 \sum_{n=0}^{\infty} \left(\frac{1}{3}\right)^n = 5 \cdot \frac{1}{1 - \frac{1}{3}} = \frac{15}{2}.$$

The case (c). We have that (with the substitution $m = n - 1$)

$$\sum_{n=1}^{\infty} \left(\frac{3}{4^{2n-1}} + \frac{2}{4^{2n}}\right) = \frac{3}{4} \sum_{n=1}^{\infty} \left(\frac{1}{4^{2n-2}}\right) + \frac{2}{16} \sum_{n=1}^{\infty} \left(\frac{1}{4^{2n-2}}\right) =$$
$$\left(\frac{3}{4} + \frac{2}{16}\right) \sum_{m=0}^{\infty} \frac{1}{4^{2m}} = \frac{14}{16} \sum_{m=0}^{\infty} \left(\frac{1}{16}\right)^m = \frac{14}{16} \cdot \frac{1}{1 - \frac{1}{16}} = \frac{14}{15}.$$

The series of linear combinations was expressed as a linear combination of series (to be more precise, as a sum of series with factoring out the constants), which is a valid modification supposing the obtained series are absolutely convergent.

The case (d). From the partial sum

$$s_n = \frac{1}{3} + \frac{2}{3^2} + \frac{3}{3^3} + \cdots + \frac{n}{3^n}, \quad n \in \mathbb{N},$$

we immediately get that

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 ρ, $\rho^{-1} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$, 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, $x \in \mathbb{C}$

$$e^{x+y} = e^x \cdot e^y,$$

see (2) and the theorem 5.4.4(3).

Substitute the values $x = it$, where $i \in \mathbb{C}$ is the imaginary unit, $t \in \mathbb{R}$ arbitrary.

$$e^{it} = 1 + it - \frac{1}{2}t^2 - i\frac{1}{3!}t^3 + \frac{1}{4!}t^4 + i\frac{1}{5!}t^5 - \cdots.$$

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 θ is the corresponding angle.
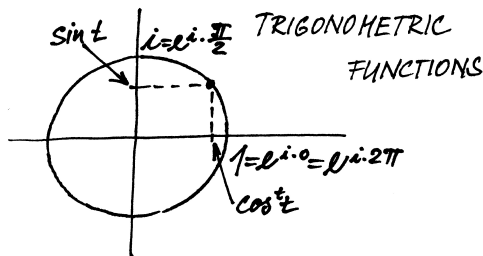
Differentiating the parametric description of the points of the circle $t \mapsto e^{it}$, gives the vectors of "velocities" which are easily computed. Differentiating the real and imaginary parts separately (assuming that the real power series can be differentiated term by term) gives :

$$\left(e^{it}\right)' = \left(1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 \ldots\right)' + i\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 \ldots\right)'$$
$$= -\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 \ldots\right) + i\left(1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 \ldots\right)$$

which means $\left(e^{it}\right)' = i \cdot e^{it}$. 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 2π stands for the

$$\frac{s_n}{3} = \frac{1}{3^2} + \frac{2}{3^3} + \cdots + \frac{n-1}{3^n} + \frac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Therefore,

$$s_n - \frac{s_n}{3} = \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \cdots + \frac{1}{3^n} - \frac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Since $\lim\limits_{n\to\infty} \frac{n}{3^{n+1}} = 0$, we get that

$$\sum_{n=1}^{\infty} \frac{n}{3^n} = \lim_{n\to\infty} \frac{3}{2}\left(s_n - \frac{s_n}{3}\right) = \frac{3}{2}\lim_{n\to\infty}\sum_{k=1}^{n}\frac{1}{3^k} =$$
$$\frac{3}{2}\sum_{k=1}^{\infty}\left(\frac{1}{3}\right)^k = \frac{3}{2}\left(\frac{1}{1-\frac{1}{3}} - 1\right) = \frac{3}{4}.$$

The case (e). It suffices to use the form (this is the so-called partial fraction decomposition)

$$\frac{1}{(3n+1)(3n+4)} = \frac{1}{3}\cdot\frac{1}{3n+1} - \frac{1}{3}\cdot\frac{1}{3n+4}, \quad n \in \mathbb{N}\cup\{0\},$$

which gives

$$\sum_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)} =$$
$$\lim_{n\to\infty}\frac{1}{3}\left(1 - \frac{1}{4} + \frac{1}{4} - \frac{1}{7} + \frac{1}{7} - \frac{1}{10} + \cdots + \frac{1}{3n+1} - \frac{1}{3n+4}\right)$$
$$= \lim_{n\to\infty}\frac{1}{3}\left(1 - \frac{1}{3n+4}\right) = \frac{1}{3}.$$

$\square$

5.H.6. Verify that

$$\sum_{n=1}^{\infty}\frac{1}{n^2} < \sum_{n=0}^{\infty}\frac{1}{2^n}.$$

Solution. We can immediately see that

$$1 \le 1, \quad \frac{1}{2^2} + \frac{1}{3^2} < 2\cdot\frac{1}{2^2} = \frac{1}{2},$$
$$\frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} < 4\cdot\frac{1}{4^2} = \frac{1}{4},$$

or, in general:

$$\frac{1}{(2^n)^2} + \cdots + \frac{1}{(2^{n+1}-1)^2} < 2^n\cdot\frac{1}{(2^n)^2} = \frac{1}{2^n}, \quad n \in \mathbb{N}.$$

Hence (by comparing the terms of both of the series) we get the wanted inequality, from which, by the way, it follows that the series $\sum_{n=1}^{\infty}\frac{1}{n^2}$ converges absolutely.

Eventually, let us specify that

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6} < 2 = \sum_{n=0}^{\infty}\frac{1}{2^n}.$$

$\square$

5.H.7. Examine convergence of the series

$$\sum_{n=1}^{\infty}\ln\frac{n+1}{n}.$$

Solution. Let us try to add up the terms of this series. We have that

$$\sum_{n=1}^{\infty}\ln\frac{n+1}{n} = \lim_{n\to\infty}\left(\ln\frac{2}{1} + \ln\frac{3}{2} + \ln\frac{4}{3} + \cdots + \ln\frac{n+1}{n}\right)$$
$$= \lim_{n\to\infty}\ln\frac{2\cdot3\cdot4\cdots(n+1)}{1\cdot2\cdot3\cdots n} = \lim_{n\to\infty}\ln(n+1) = +\infty.$$

Thus the series diverges to $+\infty$. $\square$

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* π, also called Archimedes' constant or the Ludolphian number,[15] 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 $\cos t \simeq 1 - (1/2)t^2 + (1/24)t^4 - (1/720)t^6 + (1/40320)t^8 - (1/3628800)t^{10}$. Ask Maple to find the first positive root. The result is $\pi \simeq 3.14159172323226$, for which the first 5 decimal points are correct. Compare the result 3.184900868 from the approximation of order 4.

The explicit representation of trigonometric functions in terms of the power series is now apparent:

$$\cos t = \operatorname{Re}\mathrm{e}^{it} = 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 - \frac{1}{6!}t^6 +$$
$$\cdots + (-1)^k\frac{1}{(2k)!}t^{2k} + \cdots$$
$$\sin t = \operatorname{Im}\mathrm{e}^{it} = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 +$$
$$\cdots + (-1)^k\frac{1}{(2k+1)!}t^{2k+1} + \cdots$$
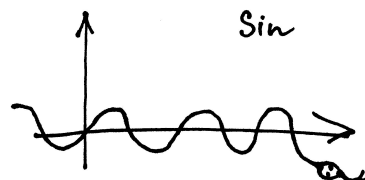
The following diagram illustrates the convergence of the series for the cosine function. It is the graph of the corresponding polynomial of degree 68. Drawing partial sums shows that the approximation near zero is very good. As the order increases, the approximation is better further away from the origin as well.
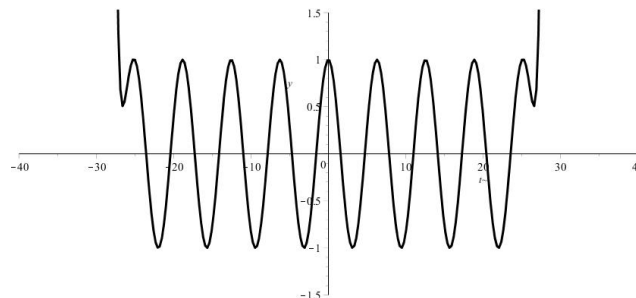
The well-known formula

$$\mathrm{e}^{it}\,\mathrm{e}^{-it} = \sin^2 t + \cos^2 t = 1$$

[15]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.

5.H.8. Prove that the series

$$\sum_{n=0}^{\infty} \operatorname{arctg} \frac{n^2+2n+3\sqrt{n}+4}{n+1}; \quad \sum_{n=1}^{\infty} \frac{3^n+1}{n^3+n^2-n}$$

do not converge.

Solution. Since

$$\lim_{n\to\infty} \operatorname{arctg} \frac{n^2+2n+3\sqrt{n}+4}{n+1} = \lim_{n\to\infty} \operatorname{arctg} \frac{n^2}{n} = \frac{\pi}{2}$$

and

$$\lim_{n\to\infty} \frac{3^n+1}{n^3+n^2-n} = \lim_{n\to\infty} \frac{3^n}{n^3} = +\infty,$$

the necessary condition $\lim_{n\to\infty} a_n = 0$ for the series $\sum_{n=n_0}^{\infty} a_n$ to converge does not hold in either case. $\square$

5.H.9. What is the series

$$\sum_{n=2}^{\infty} \frac{1}{\sqrt[n]{\ln n}}?$$

Solution. From the inequalities (consider the graph of the natural logarithm)

$$1 \le \ln n \le n, \quad n \ge 3, \ n \in \mathbb{N},$$

it follows that

$$\sqrt[n]{1} \le \sqrt[n]{\ln n} \le \sqrt[n]{n}, \quad n \ge 3, \ n \in \mathbb{N}.$$

By the squeeze theorem,

$$\lim_{n\to\infty} \sqrt[n]{\ln n} = 1, \quad \text{i. e.} \quad \lim_{n\to\infty} \frac{1}{\sqrt[n]{\ln n}} = 1.$$

Thus the series does not converge. As its terms are non-negative, it must diverge to $+\infty$. $\square$

5.H.10. Determine whether the series

(a) $\sum_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n}$;

(b) $\sum_{n=1}^{\infty} \frac{n^2+1}{n^3}$;

(c) $\sum_{n=1}^{\infty} \frac{1}{n-\ln n}$

converge.

Solution. All of the three enlisted series consist of non-negative terms only, so the series is either finite (i. e. converges), or diverges to $+\infty$. We have that

(a) $\sum_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n} \le \sum_{n=0}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{1-\frac{1}{3}} < +\infty$;

(b) $\sum_{n=1}^{\infty} \frac{n^2+1}{n^3} \ge \sum_{n=1}^{\infty} \frac{n^2}{n^3} = \sum_{n=1}^{\infty} \frac{1}{n} = +\infty$;

(c) $\sum_{n=1}^{\infty} \frac{1}{n-\ln n} \ge \sum_{n=1}^{\infty} \frac{1}{n} = +\infty$.

Hence it follows that the series (a) converges; (b) diverges to $+\infty$; (c) diverges to $+\infty$. $\square$

More interesting exercises concerning series can be found at page .

is immediate. From the derivative $(e^{it})' = i\,e^{it}$ it follows that

$$(\sin t)' = \cos t, \qquad (\cos t)' = -\sin t$$

by considering real and imaginary parts. Let t_0 denote the smallest positive number for which $e^{-it_0} = -e^{it_0}$. t_0 is the first positive zero point of the function $\cos t$. According to the definition of π, $t_0 = \frac{1}{2}\pi$.

Squaring yields $e^{i2t_0} = e^{-i2t_0} = (e^{-it_0})^2$. So π 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 2π. 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

(1) $$\cos t = \frac{1}{2}(e^{it} + e^{-it})$$

(2) $$\sin t = \frac{1}{2i}(e^{it} - e^{-it}).$$

Thus the product of these functions can be expressed as

$$\sin t \cos t = \frac{1}{4i}(e^{it} - e^{-it})(e^{it} + e^{-it})$$

$$= \frac{1}{4i}(e^{i2t} - e^{-i2t}) = \frac{1}{2}\sin 2t.$$

Further, by utilizing our knowledge of derivatives:

$$\cos 2t = (\frac{1}{2}\sin 2t)' = (\sin t \cos t)' = \cos^2 t - \sin^2 t.$$

The properties of other trigonometric functions

$$\tan t = \frac{\sin t}{\cos t}, \qquad \cot t = (\tan t)^{-1}$$

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 the red one and the green one on the left, and the red one and the green one on the right, respectively):
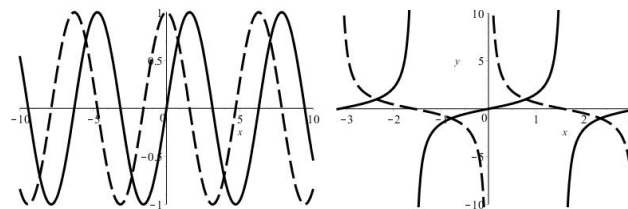
Cyclometric functions are the functions inverse to trigonometric functions. Since the trigonometric functions all have period 2π, their inverses can be defined only inside one period, and further, only on the part where the given function is either increasing or decreasing. Two inverse trigonometric functions are

$$\arcsin = \sin^{-1}$$

with domain $[-1, 1]$ and range $[-\pi/2, \pi/2]$ and

$$\arccos = \cos^{-1}$$

I. Power series

In the previous chapter, we examined whether it makes sense to assign a value to a sum of infinitely many numbers. Now we will turn our attention to the problem what sense the sum of infinitely many functions may have.

5.I.1. Determine the radius of convergence of the following power series:

i) $\sum\limits_{n=1}^{\infty} \frac{2^n}{n} x^n$

ii) $\sum\limits_{n=1}^{\infty} \frac{1}{(1+i)^n} x^n$

Solution.

i) From we get that
$$r = \frac{1}{\limsup\limits_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|} = \frac{1}{2}.$$

Thus the power series converges exactly for the real numbers $x \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ (alternatively, the complex numbers $|x| < \frac{1}{2}$). Let us notice that the series diverges for $x = \frac{1}{2}$ (it is harmonic), but on the other hand, it converges for $x = -\frac{1}{2}$ (alternating harmonic series). To determine the convergence for any x lying in the complex plane on the circle of radius $\frac{1}{2}$ is a much harder question which goes beyond our lectures.

ii)
$$r = \limsup\limits_{n\to\infty} \left| \sqrt[n]{\frac{1}{(1+i)^n}} \right| = \limsup\limits_{n\to\infty} \left| \frac{1}{1+i} \right| = \frac{\sqrt{2}}{2},$$

$\square$

5.I.2. Determine the radius r of convergence of the power series

(a) $\sum\limits_{n=1}^{\infty} \frac{(-1)^{n+1}}{n \cdot 8^n} x^n$;

(b) $\sum\limits_{n=1}^{\infty} (-4n)^n x^n$;

(c) $\sum\limits_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)^{n^2} x^n$;

(d) $\sum\limits_{n=1}^{\infty} \frac{n^5}{(2+(-1)^n)^n} x^n$.

Solution. It holds that

(a) $\lim\limits_{n\to\infty} \sqrt[n]{|a_n|} = \lim\limits_{n\to\infty} \frac{1}{\sqrt[n]{n} \cdot 8} = \frac{1}{8}$;

(b) $\lim\limits_{n\to\infty} \sqrt[n]{|a_n|} = \lim\limits_{n\to\infty} 4n = +\infty$;

(c) $\lim\limits_{n\to\infty} \sqrt[n]{|a_n|} = \lim\limits_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e$;

with domain $[-1, 1]$ and range $[0, \pi]$. See the left-hand illustration..
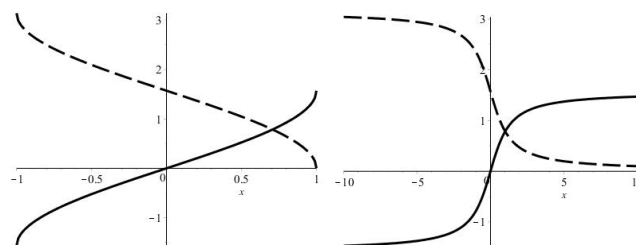
The remaining functions are (displayed in the diagram on the right)
$$\arctan = \tan^{-1}$$
with domain $\mathbb{R}$ and range $(-\pi/2, \pi/2)$, and finally
$$\operatorname{arccot} = \cot^{-1}$$
with domain $\mathbb{R}$ and range $(0, \pi)$.

The *hyperbolic functions* are also of some importance. Two basic ones are
$$\sinh x = \frac{1}{2}(e^x - e^{-x}), \qquad \cosh x = \frac{1}{2}(e^x + e^{-x}).$$
The name indicates that they should have something in common with a hyperbola. From the definition,
$$(\cosh x)^2 - (\sinh x)^2 = 2\frac{1}{2}(e^x e^{-x}) = 1.$$
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
$$\cosh x = \cos(ix), \qquad i \sinh x = \sin(ix).$$

5.4.12. 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 ρ 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 ρ 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

(d) $\limsup\limits_{n\to\infty} \sqrt[n]{|a_n|} = \limsup\limits_{n\to\infty} \frac{\sqrt[n]{n^5}}{2+(-1)^n}$

$\limsup\limits_{n\to\infty} \frac{(\sqrt[n]{n})^5}{2+(-1)^n} = 1.$

Therefore, the radius of convergence is (a) $r = 8$, (b) $r = 0$, (c) $r = 1/e$, (d) $r = 1$. $\qquad\square$

5.I.3. Calculate the radius r of convergence of the power series

$$\sum_{n=1}^{\infty} e^{in} \frac{\sqrt[3]{n^3+n}\cdot 3^n}{\sqrt[3]{n^4+2n^3+1}\cdot \pi^n} (x-2)^n.$$

Solution. 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 series

$$\sum_{n=1}^{\infty} \frac{\sqrt[3]{n^3+n}\cdot 3^n}{\sqrt[3]{n^4+2n^3+1}\cdot \pi^n} x^n.$$

Since

$$\lim_{n\to\infty} \sqrt[n]{n^a} = \left(\lim_{n\to\infty} \sqrt[n]{n}\right)^a = 1 \quad \text{for } a > 0,$$

we can move to the series

$$\sum_{n=1}^{\infty} \frac{3^n}{\pi^n} x^n$$

with the same radius of convergence $r = \pi/3$. $\qquad\square$

5.I.4. Give an example of a power series centered at the origin which, on the interval $(-3, 3)$, determines the function

$$\frac{1}{x^2-x-12}.$$

Solution. As

$$\frac{1}{x^2-x-12} = \frac{1}{(x-4)(x+3)} = \frac{1}{7}\left(\frac{1}{x-4} - \frac{1}{x+3}\right)$$

and

$$\frac{1}{x-4} = -\frac{\frac{1}{4}}{1-\frac{x}{4}} = -\frac{1}{4}\left(1 + \frac{x}{4} + \frac{x^2}{4^2} + \cdots + \frac{x^n}{4^n} + \cdots\right),$$

$$\frac{1}{x+3} = \frac{\frac{1}{3}}{1-\left(-\frac{x}{3}\right)} = \frac{1}{3}\left(1 - \frac{x}{3} + \frac{x^2}{3^2} + \cdots + \frac{(-x)^n}{3^n} + \cdots\right),$$

we get

$$\frac{1}{x^2-x-12} = -\frac{1}{28}\sum_{n=0}^{\infty} \frac{x^n}{4^n} - \frac{1}{21}\sum_{n=0}^{\infty} \frac{(-x)^n}{3^n}$$

$$= \sum_{n=0}^{\infty} \left(\frac{(-1)^{n+1}}{21\cdot 3^n} - \frac{1}{28\cdot 4^n}\right) x^n.$$

$\qquad\square$

5.I.5. Approximate the number $\sin 1°$ with error less than 10^{-10}.

Solution. We know that

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}, \quad x \in \mathbb{R}.$$

here. It is easily obtained in Maple, for instance, by the procedure "series". For illustration, here are two examples:

Begin with

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \cdots.$$

Since $e^0 = 1$, we search for a power series centered at $x = 1$ for the inverse function $\ln x$. So assume

$$\ln x = a_0 + a_1(x-1) + a_2(x-1)^2 + a_3(x-1)^3 + a_4(x-1)^4 + \cdots.$$

Apply the equality $x = e^{\ln x}$, regroup the coefficients by the powers of x and substitute. The result is:

$$x = a_0 + a_1\left(x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \cdots\right)$$

$$+ a_2\left(x + \frac{1}{2}x^2 + \cdots\right)^2 + a_3\left(x + \frac{1}{2}x^2 + \cdots\right)^3 + \cdots$$

$$= a_0 + a_1 x + \left(\frac{1}{2}a_1 + a_2\right)x^2 + \left(\frac{1}{6}a_1 + a_2 + a_3\right)x^3$$

$$+ \left(\frac{1}{24}a_1 + \left(\frac{1}{4} + \frac{2}{6}\right)a_2 + \frac{3}{2}a_3 + a_4\right)x^4 + \cdots.$$

Comparing the coefficients of the corresponding powers on both sides, gives

$$a_0 = 0, \ a_1 = 1, \ a_2 = -\frac{1}{2}, \ a_3 = \frac{1}{3}, \ a_4 = -\frac{1}{4}, \ldots.$$

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

$$\sin t = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 + \cdots$$

and the (unknown so far) power series for its inverse centered at zero (since $\sin 0 = 0$)

$$\arcsin t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + \cdots.$$

Substitution gives

$$t = a_0 + a_1\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 + \cdots\right) +$$

$$a_2\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 + \cdots\right)^2 + \cdots$$

$$= a_0 + a_1 t + a_2 t^2 + \left(-\frac{1}{6}a_1 + a_3\right)t3 +$$

$$\left(-\frac{2}{6}a_2 + a_4\right)t^4 + \left(\frac{1}{120}a_1 - \frac{3}{6}a_3 + a_5\right)t^5 + \cdots,$$

hence

$$\arcsin t = t + \frac{1}{6}t^3 + \frac{3}{40}t^5 + \cdots.$$

(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 differential equation for the coefficients a_n is easily obtained since $(x^{n+1})' = (n+1)x^n$. Therefore, from the condition that

Substituting $x = \pi/180$ gives us that the partial sums on the right side will approximate $\sin 1°$. It remains to determine the sufficient number of terms to add up in order to provably get the error below 10^{-10}. The series

$$\frac{\pi}{180} - \frac{1}{3!}\left(\frac{\pi}{180}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{180}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{180}\right)^7 + \cdots$$
$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\left(\frac{\pi}{180}\right)^{2n+1}$$

is alternating with the property that the sequence of the absolute values of its terms is decreasing. If we replace any such convergent series with its partial sum, the error we thus make will be less than the absolute value of the first term not included in the partial sum. (We do not give a proof of this theorem.) The error of the approximation

$$\sin 1° \approx \frac{\pi}{180} - \frac{\pi^3}{180^3 \cdot 3!}$$

is thus less than

$$\frac{\pi^5}{180^5 \cdot 5!} < 10^{-10}.$$

$\square$

the exponential function has its derivative equal to its value at every point,

$$a_{n+1} = \frac{1}{n+1}a_n, \qquad a_0 = 1$$

and hence it is clear that $a_n = \frac{1}{n!}$.

5.I.6. Determine the radius r of convergence of the power series $\sum\limits_{n=0}^{\infty} \frac{2^{2n} \cdot n!}{(2n)!} x^n$. $\bigcirc$

5.I.7. Calculate the radius of convergence for $\sum_{n=1}^{\infty} 2^{\sqrt{n}} x^n$. $\bigcirc$

5.I.8. Without calculation determine the radius of convergence of the power series $\sum\limits_{n=1}^{\infty} \frac{5}{n \cdot 3^{n-1}} x^{n-1}$. $\bigcirc$

5.I.9. Find the domain of convergence of the power series $\sum\limits_{n=1}^{\infty} \frac{\sqrt{n+1}}{3\sqrt{n}} x^n$. $\bigcirc$

5.I.10. Determine for which $x \in \mathbb{R}$ the power series $\sum\limits_{n=1}^{\infty} \frac{(-3)^n}{\sqrt{n^4+2n^3+111}} (x-2)^n$ converges. $\bigcirc$

5.I.11. Is the radius of convergence of the power series

$$\sum_{n=0}^{\infty} a_n x^n, \qquad \sum_{n=1}^{\infty} \frac{a_{n-1}}{n} x^n$$

common to all sequences $\{a_n\}_{n=0}^{\infty}$ of real numbers? $\bigcirc$

5.I.12. Decide whether the following implications hold:

(a) If the limit $\lim\limits_{n\to\infty} \sqrt[3n]{a_n^2}$ exists and is finite, then the power series

$$\sum_{n=1}^{\infty} a_n(x - x_0)^n$$

converges absolutely at at least two distinct points x.

(b) Conditional 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.

(c) If a series $\sum_{n=0}^{\infty} a_n$ satisfies

$$\lim_{n\to\infty} a_n^2 = 0,$$

then it is convergent.

(d) 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.I.13. Approximate $\cos\frac{\pi}{10}$ with error less than 10^{-5}.   ○

5.I.14. For the convergent series

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{\sqrt{n}+100},$$

bound the error of its approximation by the partial sum $s_{9\,999}$.

○

5.I.15. 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 $a_n x^n$.   ○

5.I.16. 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 $\mathbb{R}$?   ○

5.I.17. Does the series

$$\sum_{n=1}^{\infty} \frac{n\,x}{n^4 + x^2}, \quad \text{kde} \quad x \in \mathbb{R},$$

converge uniformly on the real line?   ○

5.I.18. 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}.

○

5.I.19. Determine the power series centered at $x_0 = 0$ of the function

$$f(x) = \int_0^x e^{t^2}\,dt, \quad x \in \mathbb{R}.$$

○

5.I.20. Using the integral test, find the values $a > 0$ for which the series

$$\sum_{n=1}^{\infty} \frac{1}{n^a}$$

converges.   ○

5.I.21. 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.I.22. Determine all $x \in \mathbb{R}$ for which the power series $\sum\limits_{i=1}^{\infty} \frac{x^{2n}}{n^2}$ is convergent. ○

5.I.23. For which $x \in \mathbb{R}$ does the series
$$\sum_{n=1}^{\infty} \frac{\ln(n!)}{n^x}$$
converge? ○

5.I.24. 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). ○

5.I.25. Calculate the series
$$\sum_{n=1}^{\infty} \frac{1}{n \cdot 3^n}$$
with the help of an appropriate power series. ○

5.I.26. For $x \in (-1, 1)$, add
$$x - 4x^2 + 9x^3 - 16x^4 + \cdots$$
○

5.I.27. Supposing $|x| < 1$, determine the series

(a) $\sum\limits_{n=1}^{\infty} \frac{1}{2n-1} x^{2n-1}$;

(b) $\sum\limits_{n=1}^{\infty} n^2 x^{n-1}$. ○

5.I.28. Calculate
$$\sum_{n=1}^{\infty} \frac{2n-1}{(-2)^{n-1}}$$
using the power series
$$\sum_{n=0}^{\infty} (-1)^n (2n+1) x^{2n}$$
for some $x \in (-1, 1)$. ○

5.I.29. For $x \in \mathbb{R}$, calculate the series
$$\sum_{n=0}^{\infty} \frac{1}{2^n \cdot n!} x^{3n+1}.$$
○

J. Additional exercises for the whole chapter

5.J.1. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(1) = 1$, $P(2) = 28$, $P(0) = 2$, $P'(0) = 1$, $P'(1) = 9$. ○

5.J.2. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(0) = 0$, $P(1) = 4$, $P(-1) = -2$, $P'(0) = 1$, $P'(1) = 7$. ○

5.J.3. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(0) = -1$, $P(1) = -1$, $P'(-1) = 10$, $P'(0) = -1$, $P'(1) = 6$. ○

5.J.4. From the definition of a limit, prove that

$$\lim_{x \to 0} \left(x^3 - 2\right) = -2.$$

○

5.J.5. From the definition of a limit, determine

$$\lim_{x \to -1} \frac{(1+x)^2 - 3}{2},$$

i. e. write the $\delta(\varepsilon)$-formula as in the previous exercise. ○

5.J.6. From the definition of a limit, show that

$$\lim_{x \to -\infty} \frac{3(x-2)^4}{2} = +\infty.$$

○

5.J.7. Determine both one-sided limits

$$\lim_{x \to 0+} \arctan\left(\frac{1}{x}\right), \qquad \lim_{x \to 0-} \arctan\left(\frac{1}{x}\right).$$

Knowing the result, decide the existence of the limit

$$\lim_{x \to 0} \arctan\left(\frac{1}{x}\right).$$

○

5.J.8. Do the following limits exist?

$$\lim_{x \to 0} \frac{\sin x}{x^3}, \qquad \lim_{x \to 0} \frac{5x^4 + 1}{x}$$

○

5.J.9. Calculate the limit

$$\lim_{x \to 0} \frac{\tan x - \sin x}{\sin^3 x}.$$

○

5.J.10. Determine

$$\lim_{x \to \pi/6} \frac{2\sin^3 x + 7\sin^2 x + 2\sin x - 3}{2\sin^3 x + 3\sin^2 x - 8\sin x + 3}.$$

○

5.J.11. For any $m, n \in \mathbb{N}$, determine

$$\lim_{x \to 1} \frac{x^m - 1}{x^n - 1}.$$

○

5.J.12. Calculate
$$\lim_{x \to +\infty} \left(\sqrt{x^2 + x} - x \right).$$

5.J.13. Determine
$$\lim_{x \to +\infty} \left(x \sqrt{1 + x^2} - x^2 \right).$$

5.J.14. Calculate
$$\lim_{x \to 0} \frac{\sqrt{2} - \sqrt{1 + \cos x}}{\sin^2 x}.$$

5.J.15. Determine
$$\lim_{x \to 0} \frac{\sin (4x)}{\sqrt{x + 1} - 1}.$$

5.J.16. Calculate
$$\lim_{x \to 0-} \frac{\sqrt{1 + \tan x} - \sqrt{1 - \tan x}}{\sin x}.$$

5.J.17. Calculate
$$\lim_{x \to -\infty} \frac{2^x + \sqrt{1 + x^2 - x^9} - 7x^5 + 44x^2}{3^x + \sqrt[5]{6x^6 + x^2} - 18x^5 - 592x^4}.$$

5.J.18. Let $\lim_{x \to -\infty} f(x) = 0$. Is it true that $\lim_{x \to -\infty}(f(x) \cdot g(x)) = 0$ for every increasing function $g : \mathbb{R} \to \mathbb{R}$?

5.J.19. Determine the limit
$$\lim_{n \to \infty} \left(\frac{n}{n + 5} \right)^{2n-1}.$$

5.J.20. Calculate
$$\lim_{x \to 0-} \frac{\sin x - x}{x^3}.$$

5.J.21. For $x > $ e, determine the sign of the derivative of the function
$$f(x) = \arctan \left(\frac{\ln x}{-1 + \ln x} \right).$$

5.J.22. Determine all local extrema of the function
$$y = x \ln^2 x$$

defined on the interval $(0, +\infty)$.

5.J.23. Is there a real number a such that the function $f(x) = ax + \sin x$ has a global minimum on the interval $[0, 2\pi]$ at $x_0 = 5\pi/4$?

5.J.24. Find the absolute minimum of the function

$$y = e\,x - \ln x, \quad x > 0$$

on its domain.

5.J.25. Determine the maximum value of the function

$$y = \sqrt[3]{3x}\,e^{-x}, \quad x \in \mathbb{R}.$$

5.J.26. Find the absolute extrema of the polynomial $p(x) = x^3 - 3x + 2$ on the interval $[-3, 2]$.

5.J.27. Let a moving object's position in time be given as follows:

$$s(t) = -(t - 3)^2 + 16, \quad t \in [0, 7],$$

where t is the time in seconds, and the position s is measured in meters. Determine

(a) the initial (i. e. at the time $t = 0\,\text{s}$) velocity of the object;
(b) the time and position at which its velocity is zero;
(c) its velocity and acceleration at time $t = 4\,\text{s}$.

Note. The object's velocity is the derivative of its position and acceleration is the derivative of its velocity.

5.J.28. From the definition of a derivative f' of a function f at the point x_0, calculate f' for $f(x) = \sqrt{x}$ at any point $x_0 > 0$.

5.J.29. Determine whether the derivative of the function

$$f(x) = x \arctan\left(\tfrac{1}{x}\right), \quad x \in \mathbb{R} \smallsetminus \{0\}, \qquad f(0) = 0$$

exists at 0.

5.J.30. Does the derivative of the function

$$y = \sin\left(\arctan\left(\left|\,12x^{21} + 11\,\right| \cdot \frac{e^{\cos(x+2) - x^3}}{-11 - x^{12}}\right)\right) + \sin(\sin(\sin(\sin x))), \quad x \in \mathbb{R}$$

at the point $x_0 = \pi^3 + 3^\pi$ exist?

5.J.31. Determine whether the derivative of the function

$$f(x) = \left(x^2 - 1\right)\sin\left(\tfrac{1}{x+1}\right), \quad x \neq -1\,(x \in \mathbb{R}), \qquad f(-1) = 0$$

at the point $x_0 = -1$ exists.

5.J.32. Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ which is continuous on the whole real axis, but does not have derivatives at the points $x_1 = 5$, $x_2 = 9$.

5.J.33. Find functions f and g which are not differentiable anywhere, yet their composition $f \circ g$ is differentiable everywhere on the real line.

5.J.34. Using basic formulae, calculate the derivative of the function

(a) $y = \left(2 - x^2\right)\cos x + 2x \sin x, \quad x \in \mathbb{R}$;
(b) $y = \sin\left(\sin x\right), \quad x \in \mathbb{R}$;
(c) $y = \sin\left(\ln\left(x^3 + 2x\right)\right), \quad x \in (0, +\infty)$;
(d) $y = \frac{1 + x - x^2}{1 - x + x^2}, \quad x \in \mathbb{R}$.

5.J.35. Determine the derivative of the function

(a) $y = \sqrt{x\sqrt{x\sqrt{x}}}, \quad x \in (0, +\infty);$

(b) $y = \ln\left|\tan\frac{x}{2}\right|, \quad x \in \mathbb{R} \setminus \{n\pi; n \in \mathbb{Z}\}.$

5.J.36. Write down the derivative of the function

$$y = \sin\left(\sin\left(\sin x\right)\right), \quad x \in \mathbb{R}.$$

5.J.37. For the function

$$f(x) = \arccos\left(\frac{1-x}{\sqrt{2}}\right) + \sqrt[3]{x^3}$$

with maximum possible domain, calculate f' on the largest subset of $\mathbb{R}$ where this derivative exists.

5.J.38. At any point $x \notin \{n\pi; n \in \mathbb{Z}\}$, determine the first derivative of the function $y = \sqrt[3]{\sin x}$.

5.J.39. For $x \in \mathbb{R}$, differentiate

$$x\sqrt{1+x^2} + e^x\left(x^2 - 2x + 2\right).$$

5.J.40. Calculate $f'(1)$ if

$$f(x) = (x-1)(x-2)^2(x-3)^3, \quad x \in \mathbb{R}.$$

5.J.41. Determine the derivative of the function

$$y = \sqrt[3]{\frac{1+x^3}{1-x^3}}, \quad |x| \neq 1 \, (x \in \mathbb{R}).$$

5.J.42. Differentiate (with respect to the real variable x)

$$x\ln^2\left(x + \sqrt{1+x^2}\right) - 2\sqrt{1+x^2}\ln\left(x + \sqrt{1+x^2}\right) + 2x$$

at all points where the derivative exists. Simplify the obtained expression.

5.J.43. Determine f' on a maximal set if $f(x) = \log_x e$.

5.J.44. Express the derivative of the product of four functions

$$[f(x)g(x)h(x)k(x)]'$$

as a sum of products of their derivatives and themselves, supposing all of these functions are differentiable.

5.J.45. Determine the derivative of the function

$$y = \frac{x^3\,(x+1)^2\,\sqrt[3]{x+2}}{(x+3)^2}$$

for $x > 0$.

5.J.46. The rainbow. Why is the rainbow circular?

Solution. In the exercise called Snell's law we explained what causes a rainbow. It is created by sunlight being refracted while entering a droplet of water. We continue with this problem, examining how the rays behave when going through the droplets. (See the illustration.) The ray dropping onto a droplet's surface at the point A "splits". Some part of the light reflects (at the angle φ_i from the normal line) and the other part refracts inside the droplet at the marked angle φ_r. The ray, inside the droplet, reflects off the droplet's surface at the point B. Since $|OA| = |OB|$, the angle of reflection equals φ_r. Part of the light refracts out of the droplet. The reflected ray then meets the droplet's surface again at the point C and refracts towards the observer at the angle φ_i from the normal line. We omit the case of the secondary rainbow arc, which occurs when the ray reflects twice inside the droplet before refracting out of it.

Write $\alpha := \measuredangle AIC$. Since $\measuredangle OAI = \varphi_i$ and $\measuredangle OAB = \varphi_r$, it follows that $\measuredangle BAI = \varphi_i - \varphi_r$. Then

$$\measuredangle BIA = \pi - (\measuredangle ABI) - (\measuredangle BAI) = \pi - (\pi - \varphi_r) - (\varphi_i - \varphi_r) = 2\varphi_r - \varphi_i$$

and moreover,

$$\alpha = 2 \cdot \measuredangle BIA = 4\varphi_r - 2\varphi_i.$$

By Snell's law,

$$\tfrac{\sin \varphi_i}{\sin \varphi_r} = n,$$

where n stands for the refractive index for water (It is assumed that the index of refraction for air is 1). Thus,

$$\varphi_r = \arcsin\left(\tfrac{\sin \varphi_i}{n}\right)$$

whence it follows that

$$(1) \qquad \alpha = 4\arcsin\left(\frac{\sin \varphi_i}{n}\right) - 2\varphi_i.$$

For the rays leaving the droplet, the value of α is different. The admissible values of α are not distributed uniformly. If R is the droplet's radius and y is the distance of the point A from the horizontal plane going through the centre of the droplet, then

$$(2) \qquad \sin \varphi_i = \frac{y}{R} \quad \text{for} \quad y \in [0, R].$$

By the huge distance of the Sun from the Earth, it is assumed that the amount of energy coming from the Sun for $y \in [a - \delta, a + \delta]$ is independent of $a \in [\delta, R - \delta]$ but depends only on the range of the values of y for sufficiently small $\delta > 0$. Thus it makes sense to analyze the function (see (1) and (2))

$$\alpha(y) = 4\arcsin\left(\tfrac{y}{nR}\right) - 2\arcsin\left(\tfrac{y}{R}\right), \qquad y \in [0, R].$$

Select an appropriate unit of length so that $R = 1$. Consider the function

$$\alpha(x) = 4\arcsin\left(\tfrac{x}{n}\right) - 2\arcsin x, \qquad x \in [0, 1].$$

The derivative is

$$\alpha'(x) = \frac{4}{n\sqrt{1 - \frac{x^2}{n^2}}} - \frac{2}{\sqrt{1 - x^2}}, \qquad x \in (0, 1)$$

The equation $\alpha'(x) = 0$ has a unique solution

$$x_0 = \sqrt{\tfrac{4 - n^2}{3}} \in (0, 1), \quad \text{if} \quad n^2 \in (1, 4).$$

Set $n = 4/3$ (which is approximately the refractive index of water). Further,

$$\alpha'(x) > 0, \quad x \in (0, x_0), \qquad \alpha'(x) < 0, \quad x \in (x_0, 1).$$

At the point

$$x_0 = \sqrt{\tfrac{4 - \left(\frac{4}{3}\right)^2}{3}} = \tfrac{2}{3}\sqrt{\tfrac{5}{3}} \doteq 0.86,$$

the function α has a global maximum

$$\alpha(x_0) = 4 \arcsin \left(\tfrac{\sqrt{5}}{2\sqrt{3}}\right) - 2 \arcsin \left(\tfrac{2\sqrt{5}}{3\sqrt{3}}\right) \doteq 0.734 \,\text{rad} \approx 42\,°.$$

It follows that the peak of the rainbow cannot be above the level of approximately $42\,°$ with regard to the observer. the values

$$\alpha(0.74) \doteq 39.4\,°, \quad \alpha(0.94) \doteq 39.2\,°, \quad \alpha(0.8) \doteq 41.2\,°, \quad \alpha(0.9) \doteq 41.5\,°.$$

suggest (α is increasing on the interval $[0, x_0]$ and decreasing on the interval $[x_0, 1]$) that more than $20\,\%$ of the values α lie in the band from around $39\,°$ to around $42\,°$, and that $10\,\%$ lie in a band thinner than $1\,°$. Furthermore, for

$$\alpha(0.84) \doteq 41.9\,°, \qquad \alpha(0.88) \doteq 41.9\,°,$$

the rays for which α is close to $42\,°$ have the greatest intensity. This is an instance of the principle of minimum deviation: the highest concentration of diffused light happens at the rays with minimum deviation since the total angle deviation of the ray equals the angle $\delta = \pi - \alpha$.

The droplets from which the rays creating the rainbow for the observer come, lie on the surface of a cone having central angle equal to $2\alpha(x_0)$. The part of this cone which is above ground then appears as the rainbow arc to the observer (see the illustration). Thus when the sun is setting, the rainbow has the shape of a semicircle. The rainbow exists only with regard to its observer – it is not anchored in space. The circular shape of the rainbow was examined as early as 1635–1637 by René Descartes. $\qquad\square$

5.J.47. 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 $l > \sqrt{d^2 + r^2}$, is tied to B at one end, passes over the pulley, and has a weight is attached to its other end. 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? See the illustration.

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. Then 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 $l - \sqrt{(d-x)^2 + r^2 - x^2}$. Hence if $f(x)$ is the distance between the weight and the ceiling, then

$$f(x) = \sqrt{r^2 - x^2} + l - \sqrt{(d-x)^2 + r^2 - x^2}.$$

The state of the system is fully determined by the value $x \in [0, r]$ (see the illustration). So it suffices to find the global maximum of the function f on the interval $[0, r]$.

The derivative is

$$f'(x) = \tfrac{-x}{\sqrt{r^2 - x^2}} - \tfrac{-(d-x)-x}{\sqrt{(d-x)^2 + r^2 - x^2}} = \tfrac{-x}{\sqrt{r^2 - x^2}} + \tfrac{d}{\sqrt{(d-x)^2 + r^2 - x^2}}, \quad x \in (0, r).$$

Square the equation $f'(x) = 0$ to obtain

$$\tfrac{x^2}{r^2 - x^2} = \tfrac{d^2}{(d-x)^2 + r^2 - x^2}.$$

and hence

$$2dx^3 - \left(2d^2 + r^2\right) x^2 + d^2 r^2 = 0, \qquad x \in (0, r).$$

One solution to this equation is $x = d$, hence the polynomial on the left side factors into

$$(x - d)\left(2dx^2 - r^2 x - dr^2\right),$$

or

$$2d\,(x - d)\left(x - \tfrac{r^2 + r\sqrt{r^2 + 8d^2}}{4d}\right)\left(x - \tfrac{r^2 - r\sqrt{r^2 + 8d^2}}{4d}\right),$$

Hence the equation $f'(x) = 0$ has three solutions. The solution $x = d$ is outside the interval $[0, r]$ since $d > r$. The solution $x = x_1 = \frac{r^2 - r\sqrt{r^2 + 8d^2}}{4d}$ is outside the interval $[0, r]$ since $x_1 < 0$. The solution

$$x = x_0 = \frac{r^2 + r\sqrt{r^2 + 8d^2}}{4d}$$

is positive, and furthermore, $x_0 < \frac{r}{4} + \frac{3r}{4} = r$, since $r < d$. Since f' is continuous on the interval $(0, r)$, it can change sign only at x_0. From the limits

$$\lim_{x \to 0+} f'(x) = \frac{d}{\sqrt{d^2 + r^2}}, \qquad \lim_{x \to r-} f'(x) = -\infty,$$

it follows that

$$f'(x) > 0, \quad x \in (0, x_0), \qquad f'(x) < 0, \quad x \in (x_0, r).$$

Thus the function f has its global maximum on the interval $[0, r]$ at x_0. $\qquad\square$

5.J.48. A nameless mail company can only transport parcels whose length does not exceed 108 inches, and for which the sum of the length and the maximal perimeter is at most 165 inches. Find the largest volume a parcel can have to be transported by this company.

Solution. Let M denote the length plus perimeter, p the perimeter, and x the parcel's length. Suppose the perimeter p is constant.

The wanted parcel has a shape such that for any $t \in (0, x)$, its cross section has constant perimeter p. (the maximal one).

The parcel is to have the greatest volume so that the cross section of a given perimeter has the greatest area possible. The largest planar figure of a given perimeter is a disc. Thus the desired parcel has the shape of a cylinder with height equal to x and radius $r = p/2\pi$.

Its volume is

$$V = \pi r^2 x = \frac{p^2 x}{4\pi}.$$

Consequently $p + x \leq M$ and $x \leq 108$. Thus we consider the parcel for which $p + x = M$. Its volume is

$$V(x) = \frac{(M-x)^2 x}{4\pi} = \frac{x^3 - 2Mx^2 + M^2 x}{4\pi} \quad \text{where} \quad x \in (0, 108].$$

Having calculated the derivative

$$V'(x) = \frac{3x^2 - 4Mx + M^2}{4\pi} = \frac{3(x-M)\left(x - \frac{M}{3}\right)}{4\pi}, \qquad x \in (0, 108),$$

we find that the function V is increasing on the interval $(0, 55] = (0, M/3]$ and decreasing on the interval $[55, 108] = [M/3, \min\{108, M\}]$. The greatest volume is thus obtained for $x = M/3$, where

$$V\left(\frac{M}{3}\right) = \frac{M^3}{27\pi} \doteq 0.011\,789\,M^3 \approx 0.867\,8\,\text{m}^3.$$

If the company also required that the parcel have the shape of a rectangular cuboid (or more generally a right prism of a given number of faces), we can repeat the previous reasoning for a given cross section of area S without specifying what the cross section looks like. Necessarily $S = kp^2$ for some $k > 0$ which is determined by the shape of the cross section. (If we change only the size of the sides of the polygon which is the cross section, then its perimeter will change by the same ratio. However, its area will change by the square of the ratio.) Thus the parcel's volume is the function

$$V(x) = Sx = kp^2 x = k(M - x)^2 x, \qquad x \in (0, 108].$$

The constant k does not affect the point of the global maximum of the function V, so the maximum is again at the point $x = M/3$. For instance, for the largest right prism having a square base, we have $p = M - x = 2M/3$, i. e. the length of the square's sides is $a = M/6$ and the volume is then

$$V = a^2 x = \frac{M^3}{6^2 \cdot 3} \doteq 0.009\,259\,M^3 \approx 0.681\,6\,\text{m}^3.$$

For a parcel in the shape of a ball (when x is the diameter), the condition $p + x \leq M$ can be expressed as $\pi x + x \leq M$, i. e. $x \leq M/(\pi + 1) < 108$. Thus for $x = M/(\pi + 1)$, the maximal volume is

$$V = \frac{4}{3}\pi\left(\frac{x}{2}\right)^3 = \frac{\pi M^3}{6(\pi+1)^3} \doteq 0.007\,370\,M^3 \approx 0.542\,6\,\text{m}^3.$$

Similarly, for a parcel in the shape of a cube (when x is the length of the cube's edges), the condition $p + x \leq M$ means $x \leq M/5 < 108$. Thus for $x = M/5$ the maximal volume

$$V = x^3 = \left(\tfrac{M}{5}\right)^3 = 0.008\,M^3 \approx 0.588\,9 \text{ m}^3.$$

The length of the edges of the cube which has the same volume as the cylinder is

$$a = \tfrac{M}{3\sqrt[3]{\pi}} \doteq 0.227\,595\,M \approx 0.953\,849 \text{ m}.$$

Its length and perimeter sum to $5a \doteq 1.138\,M$. This is more than the company's limit by around $14\,\%$. $\qquad\square$

5.J.49. A large military area (denoted by MA) has the shape of a square, and has area of 100 km^2. It is bounded along its perimeter by a narrow path. From the starting point in one corner of MA, a target point inside MA is situated 5 km along the (boundary) path and then 2 km perpendicularly to it. One can travel along the path at 5 kph, but directly through the MA at 3 kph. At what distance should you travel along the path if you wish to get there as soon as possible?

Solution. To travel x km along the path (where $x \in [0,5]$), $x/5$ hours is needed. One way through MA is then

$$\sqrt{2^2 + (5-x)^2} = \sqrt{x^2 - 10x + 29}$$

kilometers long. This takes $\sqrt{x^2 - 10x + 29}/3$ hours. Altogether, the journey takes

$$f(x) = \tfrac{1}{5}x + \tfrac{1}{3}\sqrt{x^2 - 10x + 29}$$

hours. The only zero point of the function

$$f'(x) = \tfrac{1}{5} + \tfrac{1}{3}\,\frac{x-5}{\sqrt{x^2-10x+29}}$$

is $x = 7/2$. Since the derivative f' exists at every point of the interval $[0,5]$ and since

$$f\left(\tfrac{7}{2}\right) = \tfrac{23}{15} < f(5) = \tfrac{5}{3} < f(0) = \tfrac{\sqrt{29}}{3},$$

the function f has its absolute minimum at the point $x = 7/2$ Thus you should go 3.5 km along the path. $\qquad\square$

5.J.50. 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 $\sqrt{d^2 + l^2}$ from you (see the diagram). 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 $x \in [0,l]$, and then running along the shore to the target point $[0,l]$ (see the diagram). So the trajectory consists of two line segments (or only one segment, in the case when $x = l$). The voyage to the point $[0,x]$ on the shore takes

$$\frac{\sqrt{d^2+x^2}}{v_1} \text{ hours.}$$

The final run takes

$$\frac{l-x}{v_2} \text{ hours.}$$

The total time is given by the function

$$t(x) = \frac{\sqrt{d^2+x^2}}{v_1} + \frac{l-x}{v_2}$$

on the interval $[0,l]$. 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 $x = l$.

The first derivative is

$$t'(x) = \frac{x}{v_1\sqrt{d^2+x^2}} - \frac{1}{v_2}, \qquad x \in (0,l)$$

and the second derivative is

$$t''(x) = \frac{d^2}{v_1\sqrt{(d^2+x^2)^3}}, \qquad x \in (0,l).$$

Solve the equation

$$t'(x) = 0, \qquad \text{or} \qquad \frac{x}{\sqrt{d^2+x^2}} = \frac{v_1}{v_2}.$$

Squaring and rearranging gives

$$x^2 = \left(\tfrac{v_1}{v_2}\right)^2 \left(d^2 + x^2\right).$$

$$x^2 = \frac{\left(\tfrac{v_1}{v_2}\right)^2 d^2}{1 - \left(\tfrac{v_1}{v_2}\right)^2}, \quad \text{or} \quad x = x_0 = \frac{v_1 d}{\sqrt{v_2^2 - v_1^2}}$$

If $x_0 < l$, then $t(x)$ has a global minimum at x_0 on the interval $[0, l]$ since $\lim_{x \to 0+} t'(x) < 0$, $t'(l) > 0$, and indeed $t''(x) > 0$ on the same interval.

If $x_0 \geq l$, then $t'(x) \leq 0$ on all of the interval $[0, l]$ and so the global minimum of $t(x)$ occurs at l.

In the former case, the fastest journey (in hours) takes

$$t(x_0) = \frac{\sqrt{d^2 + x_0^2}}{v_1} + \frac{l - x_0}{v_2} = \frac{d\sqrt{v_2^2 - v_1^2}}{v_1 v_2} + \frac{l}{v_2}.$$

In the latter case, the fastest journey takes

$$t(l) = \frac{\sqrt{d^2 + l^2}}{v_1} \text{ hours.}$$

Note. An alternative simpler approach for doing the calculations is to use the variable θ instead of the variable x where $x = d \tan \theta$. The fastest journey occurs when $\sin \theta = v_1/v_2$. This is a limiting case of Snell's law.

$\square$

5.J.51. A company is looking for a rectangular patch of land with sides of lengths $5a$ and b. The company wants to enclose it with a fence and then split it into 5 equal parts (each being a rectangle with sides a, b) by further fences. For which values of a, b will the area $S = 5ab$ of the patch be maximal if the total length of the used fences is $2\,400$ m?

Solution. Reformulate the statement of the problem: Maximize $5ab$ while satisfying the condition

(1) $$6b + 10a = 2,400, \quad a, b > 0.$$

The function

$$a \mapsto 5a\,\frac{2,400 - 10a}{6}$$

defined for $a \in [0, 240]$ takes its maximal value at the point $a = 120$. Hence the result is

$$a = 120 \text{ m}, \quad b = 200 \text{ m}.$$

The value of b follows immediately from (1). $\square$

5.J.52. 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.J.53. Determine the dimensions of an (open) swimming pool whose volume is 32 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. $\bigcirc$

5.J.54. Express 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. $\bigcirc$

5.J.55. With the help of the first derivative, find the real number $a > 0$ for which the sum $a + 1/a$ is minimal. Now solve this problem without using the differential calculus. $\bigcirc$

5.J.56. Inscribe a rectangle with the greatest perimeter possible into a semidisc with radius r. Determine the rectangle's perimeter. $\bigcirc$

5.J.57. Among the rectangles with perimeter $4c$, find the one having the greatest area (if such one exists) and determine the lengths of its sides. $\bigcirc$

5.J.58. Find the height h and the radius r of the largest (greatest volume) cone which fits into a ball of radius R. $\bigcirc$

5.J.59. From all triangles with given perimeter p, select the one with the greatest area. $\bigcirc$

5.J.60. A parabola is given by the equation $2x^2 - 2y = 9$. Find the points of the parabola which are closest to the origin. ◯

5.J.61. Your task is to create a one-litre tin having the "usual" shape of a cylinder so that the minimal amount of material would be used. Determine the proper ratio between its height h and radius r. ◯

5.J.62. Determine the distance from the point $[3, -1] \in \mathbb{R}^2$ to the parabola $y = x^2 - x + 1$. ◯

5.J.63. Determine the distance of the point $[-4, -2] \in \mathbb{R}^2$ from the parabola $y = x^2 + x + 1$. ◯

5.J.64. At time $t = 0$, a car left the point $A = [5, 0]$ at the speed of 4 units per second in the direction $(-1, 0)$. At the same time, another car left $B = [-2, -1]$ at the speed of 2 units per second in the direction $(0, 1)$. When will the cars be closest to each other, and what will the distance between them be at that moment? ◯

5.J.65. At the time $t = 0$, a car left the point $A = [0, 0]$ at 2 units per second in the direction $(1, 0)$. At the same time, another car left the point $B = [1, -1]$ at 3 units per second in the direction $(0, 1)$. When will they be closest to each other and what will the distance be? ◯

5.J.66. If a cone has a base of radius r and height h, then its surface area (not including the base) is $\pi r h$ and its volume is $V = \frac{1}{3}\pi r^2 h$. Determine the maximum possible volume of a cone with total surface area (including the base 3π cm^2. ◯

5.J.67. Suppose you own an excess of funds without the possibility of investing outside your own factory. This acts as a regulated market with a nearly unlimited demand and a limited access to some key raw materials, which allows you to produce at most $10\,000$ products per day. You know that the raw profit p and the expenses e, as functions of a variable x which determines the average number of products per day, satisfy

$$v(x) = 9x, \quad n(x) = x^3 - 6x^2 + 15x, \quad x \in [0, 10].$$

At what production will you profit the most from your factory? ◯

5.J.68. Determine

$$\lim_{x \to 0} \left(\cot x - \frac{1}{x} \right).$$

Solution. Notice that

$$\lim_{x \to 0+} \cot x = +\infty, \qquad \lim_{x \to 0+} \frac{1}{x} = +\infty,$$

$$\lim_{x \to 0-} \cot x = -\infty, \qquad \lim_{x \to 0-} \frac{1}{x} = -\infty,$$

we can see that both one-sided limits are of the type $\infty - \infty$. Thus we consider the (two-sided) limit.

We write the cotangent function as the ratio of the cosine and the sine and convert the fractions to a common denominator.

$$\lim_{x \to 0} \left(\cot x - \frac{1}{x} \right) = \lim_{x \to 0} \frac{x \cos x - \sin x}{x \sin x}.$$

Thus we obtain an expression of the type $0/0$ for which we get (by l'Hospital's rule)

$$\lim_{x \to 0} \frac{x \cos x - \sin x}{x \sin x} = \lim_{x \to 0} \frac{\cos x - x \sin x - \cos x}{\sin x + x \cos x} = \lim_{x \to 0} \frac{-x \sin x}{\sin x + x \cos x}.$$

By one more use of l'Hospital's rule for the type $0/0$, we get

$$\lim_{x \to 0} \frac{-x \sin x}{\sin x + x \cos x} = \lim_{x \to 0} \frac{-\sin x - x \cos x}{\cos x + \cos x - x \sin x} = \frac{0 - 0}{1 + 1 - 0} = 0.$$

□

5.J.69. Determine the limit

$$\lim_{x \to 1-} (1 - x) \tan \frac{\pi x}{2}.$$

◯

5.J.70. Calculate

$$\lim_{x \to \frac{\pi}{2}-} \left(\frac{\pi}{2} - x\tan x \right).$$

◯

5.J.71. Using l'Hospital's rule, determine

$$\lim_{x \to +\infty} \left(\left(3^{\frac{1}{x}} - 2^{\frac{1}{x}} \right) x \right).$$

◯

5.J.72. Calculate

$$\lim_{x \to 1} \left(\frac{1}{2\ln x} - \frac{1}{x^2 - 1} \right).$$

◯

5.J.73. By l'Hospital's rule, calculate the limit

$$\lim_{x \to +\infty} \left(\cos \frac{2}{x} \right)^{x^2}.$$

◯

5.J.74. Determine

$$\lim_{x \to 0} \left(1 - \cos x \right)^{\sin x} = \dots$$

◯

5.J.75. Determine the following limits

$$\lim_{x \to 0+} x^{\frac{\alpha}{\ln x}}, \qquad \lim_{x \to +\infty} x^{\frac{\alpha}{\ln x}},$$

where $\alpha \in \mathbb{R}$ is arbitrary.

◯

5.J.76. By any means, verify that

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1.$$

◯

5.J.77. By applying the ratio test (also called D'Alembert's criterion; see 5.4.5), determine whether the infinite series

(a) $\sum_{n=1}^{\infty} \frac{2^n \cdot (n+1)^3}{3^n}$;

(b) $\sum_{n=1}^{\infty} \frac{6^n}{n!}$;

(c) $\sum_{n=1}^{\infty} \frac{n^n}{n^2 \cdot n!}$

converge.

Solution. Since ($a_n \geq 0$ for all n)

(a) $\lim_{n \to \infty} \frac{a_{n+1}}{a_n} = \lim_{n \to \infty} \frac{2^{n+1} \cdot (n+2)^3 \cdot 3^n}{3^{n+1} \cdot 2^n \cdot (n+1)^3} = \lim_{n \to \infty} \frac{2(n+2)^3}{3(n+1)^3} = \lim_{n \to \infty} \frac{2n^3}{3n^3} = \frac{2}{3} < 1$;

(b) $\lim_{n \to \infty} \frac{a_{n+1}}{a_n} = \lim_{n \to \infty} \left(\frac{6^{n+1}}{(n+1)!} \cdot \frac{n!}{6^n} \right) = \lim_{n \to \infty} \frac{6}{n+1} = 0 < 1$;

(c) $\lim_{n \to \infty} \frac{a_{n+1}}{a_n} = \lim_{n \to \infty} \left(\frac{(n+1)^{n+1}}{(n+1)^2 \cdot (n+1)!} \cdot \frac{n^2 \cdot n!}{n^n} \right) = \lim_{n \to \infty} \frac{n^2}{(n+1)^2} \cdot \lim_{n \to \infty} \frac{(n+1)^n}{n^n} = \lim_{n \to \infty} \frac{n^2}{n^2} \cdot \lim_{n \to \infty} \left(1 + \frac{1}{n} \right)^n = 1 \cdot e > 1$,

the series (a) converges; (b) converges; (c) does not converge (it diverges to $+\infty$). □

5.J.78. By applying the root test (Cauchy's criterion), determine whether or not the infinite series

(a) $\sum\limits_{n=1}^{\infty} \frac{1}{\ln^n (n+1)}$;

(b) $\sum\limits_{n=1}^{\infty} \frac{\left(\frac{n+1}{n}\right)^{n^2}}{n^3 \cdot 3^n}$;

(c) $\sum\limits_{n=1}^{\infty} \arcsin^n \frac{2n}{2^n}$

converge.

Solution. Consider the series with non-negative terms only, where

(a) $\lim\limits_{n\to\infty} \sqrt[n]{a_n} = \lim\limits_{n\to\infty} \frac{1}{\ln(n+1)} = 0 < 1$;

(b) $\lim\limits_{n\to\infty} \sqrt[n]{a_n} = \lim\limits_{n\to\infty} \frac{\left(\frac{n+1}{n}\right)^n}{\sqrt[n]{n^3}\cdot 3} = \frac{\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n}{3\left(\lim\limits_{n\to\infty}\sqrt[n]{n}\right)^3} = \frac{e}{3} < 1$;

(c) $\lim\limits_{n\to\infty} \sqrt[n]{a_n} = \lim\limits_{n\to\infty} \arcsin \frac{2n}{2^n} = \arcsin 0 = 0 < 1$.

So each of the above series converges. □

5.J.79. Determine whether or not the series

(a) $\sum\limits_{n=1}^{\infty} (-1)^n \ln\left(1+\frac{1}{2^n}\right)$;

(b) $\sum\limits_{n=1}^{\infty} \frac{(-2)^{n^2}}{n!}$;

(c) $\sum\limits_{n=1}^{\infty} \frac{(-3)^n}{(6+(-1)^n)^n}$

converge.

Solution. The case (a). By l'Hospital's rule,

$$\lim_{x\to+\infty} \frac{\ln\left(1+\frac{1}{2^x}\right)}{\frac{1}{2^x}} = \lim_{x\to+\infty} \frac{\frac{1}{1+\frac{1}{2^x}}\left(1+\frac{1}{2^x}\right)'}{\left(\frac{1}{2^x}\right)'} = \lim_{x\to+\infty} \frac{1}{1+\frac{1}{2^x}} = 1,$$

hence

$$0 < \ln\left(1+\frac{1}{2^n}\right) \le \frac{2}{2^n}$$

for all sufficiently large $n \in \mathbb{N}$. However, the series $\sum_{n=1}^{\infty} \frac{2}{2^n}$ is convergent. So

$$\sum_{n=1}^{\infty} \ln\left(1+\frac{1}{2^n}\right) < +\infty.$$

The series converges (absolutely).

The case (b). The ratio test gives

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = \lim_{n\to\infty} \frac{2^{(n+1)^2}\cdot n!}{(n+1)!\cdot 2^{n^2}} = \lim_{n\to\infty} \frac{2^{2n+1}}{n+1} = \lim_{n\to\infty} \frac{2\cdot 4^n}{n+1} = +\infty.$$

Thus the series does not converge.

The case (c). Use the general version of the root test

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} = \limsup_{n\to\infty} \frac{3}{6+(-1)^n} = \frac{3}{5} < 1.$$

It follows that the series is (absolutely) convergent. □

5.J.80. Determine whether or not the following alternating series converge:

(a) $\sum\limits_{n=1}^{\infty} (-1)^n \frac{n^2+3n-1}{(3n-2)^2}$;

(b) $\sum\limits_{n=1}^{\infty} (-1)^{n-1} \frac{3n^4-3n^3+9n-1}{(5n^3-2)\cdot 4^n}$.

Solution. Case (a). Since

$$\lim_{n\to\infty} \frac{n^2+3n-1}{(3n-2)^2} = \lim_{n\to\infty} \frac{n^2}{9n^2} = \frac{1}{9} \neq 0,$$

it follows that the limit

$$\lim_{n\to\infty} (-1)^n \frac{n^2+3n-1}{(3n-2)^2}$$

does not exist. Therefore, the series does not converge since a necessary condition for the convergence is not satisfied.

The case (b). When applying the ratio (or root) test, the polynomials in the numerator or in the denominator do not affect the value of the considered limit. Consider the series

$$\sum_{n=1}^{\infty} (-1)^{n-1} \tfrac{1}{4^n}$$

for which

$$\lim_{n\to\infty} \left| \tfrac{a_{n+1}}{a_n} \right| = \tfrac{1}{4} < 1.$$

This means that the original series is also (absolutely) convergent. $\square$

5.J.81. Does the following series converge?

$$\sum_{n=1}^{\infty} (-1)^{n+1} \arctan \tfrac{2}{\sqrt{3n}}$$

Solution. The sequence $\left\{2/\sqrt{3n}\right\}_{n\in\mathbb{N}}$ is decreasing and the function $y = \arctan x$ increasing (on the whole real axis). So the sequence $\left\{\arctan\left(2/\sqrt{3n}\right)\right\}_{n\in\mathbb{N}}$ is decreasing. Thus it is an alternating series such that the sequence of the absolute values of its terms is decreasing. Such an alternating series converges if and only if the sequence of its terms converges to zero (the Leibniz criterion), and this is satisfied:

$$\lim_{n\to\infty} \arctan \tfrac{2}{\sqrt{3n}} = \arctan 0 = 0, \text{ i. e. } \lim_{n\to\infty} \left((-1)^{n+1} \arctan \tfrac{2}{\sqrt{3n}} \right) = 0.$$

$\square$

5.J.82. Determine whether the series

(a) $\sum\limits_{n=1}^{\infty} \tfrac{\sin n}{n^2}$;

(b) $\sum\limits_{n=1}^{\infty} \tfrac{\cos(\pi n)}{\sqrt[3]{n^2}}$

converges absolutely, converges conditionally, or does not converge at all.

Solution. The case (a). This series converges absolutely. For instance,

$$\sum_{n=1}^{\infty} \left| \tfrac{\sin n}{n^2} \right| \leq \sum_{n=1}^{\infty} \tfrac{1}{n^2} < \sum_{n=0}^{\infty} \tfrac{1}{2^n} = 2,$$

and the second inequality is already proven.

The case (b). $\cos(\pi n) = (-1)^n$, $n \in \mathbb{N}$. So it is an alternating series such that the sequence of the absolute values of its terms is decreasing. Therefore, from the limit

$$\lim_{n\to\infty} \tfrac{1}{\sqrt[3]{n^2}} = 0$$

it follows that the series is convergent. On the other hand,

$$\sum_{n=1}^{\infty} \left| \tfrac{\cos(\pi n)}{\sqrt[3]{n^2}} \right| = \sum_{n=1}^{\infty} \tfrac{1}{\sqrt[3]{n^2}} \geq \sum_{n=1}^{\infty} \tfrac{1}{n} = +\infty.$$

The series converges conditionally. $\square$

5.J.83. Calculate the series

(a) $\sum\limits_{n=1}^{\infty} \left(\tfrac{1}{\sqrt{n}} - \tfrac{1}{\sqrt{n+1}} \right)$;

(b) $\sum\limits_{n=0}^{\infty} \tfrac{5}{3^n}$;

(c) $\sum\limits_{n=1}^{\infty} \left(\tfrac{3}{4^{2n-1}} + \tfrac{2}{4^{2n}} \right)$;

(d) $\sum\limits_{n=1}^{\infty} \tfrac{n}{3^n}$;

(e) $\sum\limits_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)}$.

Solution. The case (a). By the definition,

$$\sum_{n=1}^{\infty} \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right) =$$
$$\lim_{n \to \infty} \left(\left(\frac{1}{\sqrt{1}} - \frac{1}{\sqrt{2}} \right) + \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}} \right) + \cdots + \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right) \right) =$$
$$\lim_{n \to \infty} \left(1 + \left(-\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \right) + \cdots + \left(-\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right) - \frac{1}{\sqrt{n+1}} \right) = 1.$$

The case (b). This is a convergent geometric series with the common ratio $q = 1/3$, hence

$$\sum_{n=0}^{\infty} \frac{5}{3^n} = 5 \sum_{n=0}^{\infty} \left(\frac{1}{3} \right)^n = 5 \cdot \frac{1}{1 - \frac{1}{3}} = \frac{15}{2}.$$

The case (c). By substituting $m = n - 1$,

$$\sum_{n=1}^{\infty} \left(\frac{3}{4^{2n-1}} + \frac{2}{4^{2n}} \right) = \frac{3}{4} \sum_{n=1}^{\infty} \left(\frac{1}{4^{2n-2}} \right) + \frac{2}{16} \sum_{n=1}^{\infty} \left(\frac{1}{4^{2n-2}} \right) =$$
$$\left(\frac{3}{4} + \frac{2}{16} \right) \sum_{m=0}^{\infty} \frac{1}{4^{2m}} = \frac{14}{16} \sum_{m=0}^{\infty} \left(\frac{1}{16} \right)^m = \frac{14}{16} \cdot \frac{1}{1 - \frac{1}{16}} = \frac{14}{15}.$$

The series of linear combinations was expressed as a linear combination of a series (to be more precise, as a sum of series factoring out the constants). This is a valid modification supposing the obtained series are absolutely convergent.

The case (d). From the partial sum

$$s_n = \frac{1}{3} + \frac{2}{3^2} + \frac{3}{3^3} + \cdots + \frac{n}{3^n}, \quad n \in \mathbb{N},$$

we obtain

$$\frac{s_n}{3} = \frac{1}{3^2} + \frac{2}{3^3} + \cdots + \frac{n-1}{3^n} + \frac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Thus

$$s_n - \frac{s_n}{3} = \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \cdots + \frac{1}{3^n} - \frac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Since $\lim\limits_{n \to \infty} \frac{n}{3^{n+1}} = 0$,

$$\sum_{n=1}^{\infty} \frac{n}{3^n} = \lim_{n \to \infty} \frac{3}{2} \left(s_n - \frac{s_n}{3} \right) = \frac{3}{2} \lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{3^k} =$$
$$\frac{3}{2} \sum_{k=1}^{\infty} \left(\frac{1}{3} \right)^k = \frac{3}{2} \left(\frac{1}{1 - \frac{1}{3}} - 1 \right) = \frac{3}{4}.$$

The case (e). It suffices to use partial fraction decomposition

$$\frac{1}{(3n+1)(3n+4)} = \frac{1}{3} \cdot \frac{1}{3n+1} - \frac{1}{3} \cdot \frac{1}{3n+4}, \quad n \in \mathbb{N} \cup \{0\},$$

which gives

$$\sum_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)} = \lim_{n \to \infty} \frac{1}{3} \left(1 - \frac{1}{4} + \frac{1}{4} - \frac{1}{7} + \frac{1}{7} - \frac{1}{10} + \cdots + \frac{1}{3n+1} - \frac{1}{3n+4} \right)$$
$$= \lim_{n \to \infty} \frac{1}{3} \left(1 - \frac{1}{3n+4} \right) = \frac{1}{3}.$$

$\square$

5.J.84. Verify that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} < \sum_{n=0}^{\infty} \frac{1}{2^n}.$$

Solution.

$$1 \le 1, \quad \frac{1}{2^2} + \frac{1}{3^2} < 2 \cdot \frac{1}{2^2} = \frac{1}{2}, \quad \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} < 4 \cdot \frac{1}{4^2} = \frac{1}{4},$$

or the general bound

$$\frac{1}{(2^n)^2} + \cdots + \frac{1}{(2^{n+1}-1)^2} < 2^n \cdot \frac{1}{(2^n)^2} = \frac{1}{2^n}, \quad n \in \mathbb{N}.$$

By comparing the terms of both series, we get the desired inequality. It follows that the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is absolutely convergent.

Note that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < 2 = \sum_{n=0}^{\infty} \frac{1}{2^n}.$$

$\square$

5.J.85. Examine the convergence of the series

$$\sum_{n=1}^{\infty} \ln \frac{n+1}{n}.$$

Solution. Add the terms of this series.

$$\sum_{n=1}^{\infty} \ln \frac{n+1}{n} = \lim_{n \to \infty} \left(\ln \frac{2}{1} + \ln \frac{3}{2} + \ln \frac{4}{3} + \cdots + \ln \frac{n+1}{n} \right) =$$
$$\lim_{n \to \infty} \ln \frac{2 \cdot 3 \cdot 4 \cdots (n+1)}{1 \cdot 2 \cdot 3 \cdots n} = \lim_{n \to \infty} \ln (n+1) = +\infty.$$

Thus the series diverges to $+\infty$. $\square$

5.J.86. Prove that the series

$$\sum_{n=0}^{\infty} \arctan \frac{n^2 + 2n + 3\sqrt{n} + 4}{n+1}; \quad \sum_{n=1}^{\infty} \frac{3^n + 1}{n^3 + n^2 - n}$$

do not converge.

Solution. Since

$$\lim_{n \to \infty} \arctan \frac{n^2 + 2n + 3\sqrt{n} + 4}{n+1} = \lim_{n \to \infty} \arctan \frac{n^2}{n} = \frac{\pi}{2}$$

and

$$\lim_{n \to \infty} \frac{3^n + 1}{n^3 + n^2 - n} = \lim_{n \to \infty} \frac{3^n}{n^3} = +\infty,$$

the necessary condition $\lim_{n \to \infty} a_n = 0$ for the series $\sum_{n=n_0}^{\infty} a_n$ to converge does not hold. $\square$

5.J.87. Determine whether or not the series

$$\sum_{n=2}^{\infty} \frac{1}{\sqrt[n]{\ln n}}?$$

converges.

Solution. From the inequalities (consider the graph of the natural logarithm)

$$1 \leq \ln n \leq n, \quad n \geq 3, \ n \in \mathbb{N}$$

it follows that

$$\sqrt[n]{1} \leq \sqrt[n]{\ln n} \leq \sqrt[n]{n}, \quad n \geq 3, \ n \in \mathbb{N}.$$

By the squeeze theorem (5.2.12),

$$\lim_{n \to \infty} \sqrt[n]{\ln n} = 1, \quad \text{i. e.} \quad \lim_{n \to \infty} \frac{1}{\sqrt[n]{\ln n}} = 1.$$

Thus the series is not convergent. Since all its terms are non-negative, it diverges to $+\infty$. $\square$

5.J.88. Determine whether or not the series

(a) $\sum_{n=0}^{\infty} \frac{1}{(n+1) \cdot 3^n}$;

(b) $\sum_{n=1}^{\infty} \frac{n^2 + 1}{n^3}$;

(c) $\sum_{n=1}^{\infty} \frac{1}{n - \ln n}$

converges.

Solution. All of the three enlisted series consist of non-negative terms only. So the series either converge, or diverge to $+\infty$.

(a) $\sum\limits_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n} \leq \sum\limits_{n=0}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{1-\frac{1}{3}} < +\infty;$

(b) $\sum\limits_{n=1}^{\infty} \frac{n^2+1}{n^3} \geq \sum\limits_{n=1}^{\infty} \frac{n^2}{n^3} = \sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty;$

(c) $\sum\limits_{n=1}^{\infty} \frac{1}{n-\ln n} \geq \sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty.$

It follows that (a) converges; (b) diverges to $+\infty$; (c) diverges to $+\infty$. □

5.J.89. Begin with a square with sides of length $a > 0$. Construct a sequence of squares, each of which has as vertices the midpoints of the preceding square. Determine the sum of the areas and the sum of the perimeters of all these (infinitely many) squares. ◯

5.J.90. Let a sequence of rows of semidiscs be given, such that for each $n \in \mathbb{N}$, the n-th row contains 2^n semidiscs, each with radius of 2^{-n}. What is the area of an arbitrary figure consisting of all these semidiscs, supposing the semicircles do not overlap? ◯

5.J.91. Solve the equation

$$1 - \tan x + \tan^2 x - \tan^3 x + \tan^4 x - \tan^5 x + \cdots = \frac{\tan 2x}{\tan 2x + 1}.$$

◯

5.J.92. Determine

$$\sum\limits_{n=1}^{\infty} \left(\frac{1}{2^{n-1}} + \frac{2}{3^{n-1}}\right).$$

◯

5.J.93. Calculate

$$\sum\limits_{n=1}^{\infty} \sqrt[5n]{n^2 + 2n + 1}.$$

◯

5.J.94. Prove the convergence of the series

$$\sum\limits_{n=1}^{\infty} \frac{3^n + 2^n}{6^n}.$$

and find its value. ◯

5.J.95. Calculate the series

(a) $\sum\limits_{n=1}^{\infty} \frac{2n-1}{2^n};$

(b) $\sum\limits_{n=0}^{\infty} \frac{n+1}{3^n}.$

◯

5.J.96. Sum the series

$$\frac{1}{1\cdot 3} + \frac{1}{3\cdot 5} + \frac{1}{5\cdot 7} + \cdots = \sum\limits_{n=1}^{\infty} \frac{1}{(2n-1)(2n+1)}.$$

◯

5.J.97. Using the partial fraction decomposition, calculate

(a) $\sum\limits_{n=2}^{\infty} \frac{1}{n^2-1};$

(b) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^3+3n^2+2n}$.

○

5.J.98. Determine the value of the convergent series
$$\sum_{n=0}^{\infty} \frac{1}{4n^2-1}.$$

○

5.J.99. Calculate the series
$$\sum_{n=1}^{\infty} \frac{1}{n^2+3n}.$$

○

5.J.100. In terms of
$$s := \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \tfrac{1}{5} - \tfrac{1}{6} + \tfrac{1}{7} - \tfrac{1}{8} + \cdots,$$
express the following two series
$$\left(1 - \tfrac{1}{2} - \tfrac{1}{4}\right) + \left(\tfrac{1}{3} - \tfrac{1}{6} - \tfrac{1}{8}\right) + \cdots;$$
$$\left(1 + \tfrac{1}{3} - \tfrac{1}{2}\right) + \left(\tfrac{1}{5} + \tfrac{1}{7} - \tfrac{1}{4}\right) + \cdots$$
(both the series contain the same elements as the first one, only in a different order).

○

5.J.101. Determine whether the series
$$\sum_{n=0}^{\infty} \frac{2^n + (-2)^n}{5^n}$$
converges.

○

5.J.102. Prove the following statement: If a series $\sum_{n=0}^{\infty} a_n$ converges, then $\lim_{n\to\infty} \sin\left(3a_n + \pi\right) = 0$.

○

5.J.103. For which $\alpha \in \mathbb{R}$; $\quad \beta \in \mathbb{Z}$; $\quad \gamma \in \mathbb{R} \smallsetminus \{0\}$ do the series $\displaystyle\sum_{n=120}^{\infty} \frac{e^{-\alpha n}}{n}$; $\quad \displaystyle\sum_{n=240}^{\infty} \frac{\beta^n \cdot n!}{n^n}$; $\quad \displaystyle\sum_{n=360}^{\infty} \frac{n}{\gamma^n}$ converge?

○

5.J.104. 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.

○

5.J.105. Determine whether or not the limit
$$\lim_{n\to\infty} \left(\tfrac{1}{n^2} + \tfrac{2}{n^2} + \cdots + \tfrac{n-1}{n^2}\right)$$
is finite. Notice that one cannot use the sums
$$\sum_{n=1}^{\infty} \tfrac{1}{n^2} = \tfrac{\pi^2}{6}, \quad \sum_{n=2}^{\infty} \tfrac{n-1}{n^2} = +\infty.$$

○

5.J.106. Find all real numbers $A \geq 0$ for which the series
$$\sum_{n=1}^{\infty} (-1)^n \ln\left(1 + A^{2n}\right)$$
is convergent.

○

5.J.107. Recall that the harmonic series diverges.

$$\sum_{n=1}^{\infty} \tfrac{1}{n} = +\infty.$$

Determine whether or not the series

$$\tfrac{1}{1} + \cdots + \tfrac{1}{9} + \tfrac{1}{11} + \cdots + \tfrac{1}{19} + \tfrac{1}{21} + \cdots + \tfrac{1}{29} + \cdots$$
$$\cdots + \tfrac{1}{91} + \cdots + \tfrac{1}{99} + \tfrac{1}{111} + \cdots + \tfrac{1}{119} + \tfrac{1}{121} + \cdots$$

is also divergent. ○

5.J.108. 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.J.109. Determine whether the two series

$$\sum_{n=1}^{\infty} (-1)^n \tfrac{(n!)^2}{(2n)!}; \quad \sum_{n=1}^{\infty} (-1)^n \tfrac{n^7 - n^4 + n}{n^8 + 2n^6 + n}$$

converge absolutely, converge conditionally, or do not converge at all. ○

5.J.110. Does the series

$$\sum_{n=1}^{\infty} (-1)^{n+1} \tfrac{\sqrt[3]{n} + \sqrt[5]{n} + 1}{n + \sqrt[9]{n}}$$

converge? ○

5.J.111. Find the values of the parameter $p \in \mathbb{R}$ for which the series

$$\sum_{n=1}^{\infty} (-1)^n \sin^n \tfrac{p}{n}$$

converges. ○

5.J.112. Determine the maximal subset of $\mathbb{R}$ where the function

$$y = \operatorname{arctg}\left(x^{21} + \sin x\right) \cdot \frac{e^{\cos\left(\sqrt[5]{x} - 21 + \cos x\right) + x - 256x^3} - 11}{2 + x^{252}}$$

can be defined. ○

5.J.113. Write the maximal domain of the function $y = \frac{\arccos (\ln x)}{\sqrt{x^2 - 1}}$. ○

5.J.114. Determine the domain and the range of the function

$$y = \tfrac{x-1}{2-3x}.$$

Then determine the inverse function. ○

5.J.115. Is the function

(a) $y = \frac{\cos x}{x^3}$;
(b) $y = \frac{\cos x}{x^3} + 1$;
(c) $y = \frac{\cos x}{x^4}$;
(d) $y = \frac{\cos x}{x^4} + 1$;
(e) $y = \sin x + \tan \frac{x}{2}$;
(f) $y = \ln \frac{1+x}{1-x}$;
(g) $y = \sinh x = \frac{e^x - e^{-x}}{2}$;
(h) $y = \cosh x = \frac{e^x + e^{-x}}{2}$

with the maximal domain odd, even, or neither? ○

5.J.116. Is the function

(a) $y = \frac{\cos x}{x^3}$;
(b) $y = \frac{\cos x}{x^3} + 1$;
(c) $y = \frac{\cos x}{x^4}$;
(d) $y = \frac{\cos x}{x^4} + 1$;

(e) $y = \sin x + \tan \frac{x}{2}$;

(f) $y = \ln \frac{1+x}{1-x}$;

(g) $y = \sinh x = \frac{e^x - e^{-x}}{2}$;

(h) $y = \cosh x = \frac{e^x + e^{-x}}{2}$

with the maximal domain even, odd or neither? ◯

5.J.117. Determine whether the function

(a) $y = \sin x \cdot \ln |x|$;

(b) $y = \operatorname{arccotg} x$;

(c) $y = x^8 - \sqrt[5]{3}x^6 + 3x^2 - 6$;

(d) $y = \cos (\pi - x)$;

(e) $y = \frac{\tan x + x}{3 + 7\cos x}$

with the maximal domain is odd or even. ◯

5.J.118. Is the function

(a) $y = \ln (\cos x)$;

(b) $y = \tan (3x) + 2\sin (6x)$

with maximal domain periodic? ◯

5.J.119. Draw the graphs of the functions $f(x) = e^{|x|}$, $x \in \mathbb{R}$, $g(x) = \ln |x|$, $x \in \mathbb{R} \smallsetminus \{0\}$. ◯

5.J.120. Draw the graph of the function $y = 2^{-|x|}$, $x \in \mathbb{R}$. ◯

5.J.121. The functions

$$\sinh x = \frac{e^x - e^{-x}}{2}, \ x \in \mathbb{R}; \quad \cosh x = \frac{e^x + e^{-x}}{2}, \ x \in \mathbb{R};$$

$$\tanh x = \frac{\sinh x}{\cosh x}, \ x \in \mathbb{R}; \quad \coth x = \frac{\cosh x}{\sinh x}, \ x \in \mathbb{R} \smallsetminus \{0\}$$

are called hyperbolic functions. Determine the derivatives of these functions on their domains. ◯

5.J.122. At any point $x \in \mathbb{R}$, calculate the derivative of the area hyperbolic sine (denoted arsinh), the function inverse to the hyperbolic sine $y = \sinh x$ on $\mathbb{R}$. ◯

Remark. Note, that the inverse functions to the hyperbolic functions $y = \cosh x$, $x \in [0, +\infty)$, $y = \tanh x$, $x \in \mathbb{R}$ and $y = \coth x$, $x \in (-\infty, 0) \cup (0, +\infty)$ are called area hyperbolic functions ($y = \operatorname{arsinh} x$ belongs to them, too). They are denoted arcosh, artanh, arcoth, respectively and are defined for $x \in [1, +\infty)$, $x \in (-1, 1)$, and $x \in (-\infty, -1) \cup (1, +\infty)$, respectively. Let us add that

$$(\operatorname{arcosh} x)' = \frac{1}{\sqrt{x^2 - 1}}, \quad x > 1,$$

$$(\operatorname{artanh} x)' = \frac{1}{1 - x^2}, \quad |x| < 1,$$

$$(\operatorname{arcoth} x)' = \frac{1}{1 - x^2}, \quad |x| > 1.$$

5.J.123. Calculate the sum of the series:

$$2 + 1 + \frac{2}{2!} + \frac{1}{3!} + \frac{2}{4!} + \frac{1}{5!} + \frac{2}{6!} + \cdots$$

Solutions to the exercises

5.A.7. $P(x) = (-\frac{3}{5} - \frac{4}{5}i)x^2 + (2 + 3i)x - \frac{3}{5} - \frac{14}{5}i$.

5.A.15. Sought spline differs from the one in 5.A.14 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, the wanted spline is the function

$$S(x) = \begin{cases} -3x^3 - 4x^2 + 1 & \text{pro } x \in [-1, 0], \\ 3x^3 - 4x^2 + 1 & \text{pro } x \in [0, 1]. \end{cases}$$

5.A.16. $3x^2 - 2x - 4$.

5.A.17. $\left(2x^2 - 5\right)/3$; eg. $\left(\frac{2}{3}x^2 - \frac{5}{3}\right)^3$.

5.A.18. $a = 1$, $b = -2$, $c = 0$, $d = 1$.

5.A.19. $x^3 + x^2 - x + 2$.

5.A.20. Infinitely many.

5.A.21. $P(x) = x^3 - 2x^2 + 5x - 3$; $Q(x) = x^3 - 2x^2 + 3x - 3$.

5.A.22. $x^5 - 2x^4 - 5x + 2$.

5.A.23. x^2.

5.A.24. $x^3 - 2x + 5$; $x^3 - x + 6$.

5.A.25. Infinitely many.

5.A.26. Eg. $x^2 - 3x + 6$.

5.A.27. $S_1(x) = \frac{1}{2}(x + 1)^3 - \frac{3}{2}(x + 1) + 1$, $x \in [-1, 0]$; $S_2(x) = -\frac{1}{2}x^3 + \frac{3}{2}x^2$, $x \in [0, 1]$.

5.A.28. $S_1(x) = \frac{1}{2}(x + 1)^3 - \frac{3}{2}(x + 1) + 1$, $x \in [-1, 0]$; $S_2(x) = -\frac{1}{2}x^3 + \frac{3}{2}x^2$, $x \in [0, 1]$.

5.A.29. $S_1(x) \equiv x$; $S_2(x) \equiv x$.

5.A.30. $S_1(x) \equiv 1$; $S_2(x) \equiv 1$.

5.A.31. $S_i(x) = x + 3$, $x \in [-3 + i - 1, -3 + i]$; $i \in \{1, 2\}$.

5.A.32. $S_1(x) = 1 - \frac{11}{20}x + \frac{1}{20}x^3$; $S_2(x) = \frac{1}{2} - \frac{2}{5}(x - 1) + \frac{3}{20}(x - 1)^2 - \frac{1}{40}(x - 1)^3$.

5.B.2.

$$\sup A = 6, \qquad \inf A = -3;$$
$$\sup B = \frac{1}{4}, \qquad \inf B = -1;$$
$$\sup C = 9, \qquad \inf C = -9.$$

5.B.3. It can easily be shown that

$$\sup A = \frac{3}{2}, \qquad \inf A = 0.$$

5.B.4. Clearly

$$\inf \mathbb{N} = 1, \quad \sup \mathcal{M} = 0, \quad \inf \mathcal{J} = 0, \quad \sup \mathcal{J} = 5.$$

5.B.5. We can, for instance, set

$$M := \mathbb{Z} \setminus \mathbb{N}; \qquad N := \mathbb{N}.$$

5.B.6. Consider any singleton (one-element set) $X \subset \mathbb{R}$.

5.B.7. The set C must be a singleton. Thus, let us choose $C = \{0\}$, for example. Now we can take $A = (-1, 0)$, $B = (0, 1)$.

5.C.5. We have

$$\lim_{n \to \infty} \left(\frac{1}{n^2} + \frac{2}{n^2} + \cdots + \frac{n - 2}{n^2} + \frac{n - 1}{n^2} \right) = \lim_{n \to \infty} \left(\frac{1 + n - 1}{n^2} \cdot \frac{n - 1}{2} \right) = \frac{1}{2}.$$

5.C.6. It can easily be shown that

$$\lim_{n\to\infty} \frac{\sqrt{n^3 - 11n^2 + 2} + \sqrt[5]{n^7 - 2n^5 - n^3} - n + \sin^2 n}{2 - \sqrt[3]{5n^4 + 2n^3 + 5}} = -\infty.$$

5.C.7. The limit is equal to 1.

5.C.8. We can, for instance, set

$$x_n := n, \quad y_n := -n + 1, \qquad n \in \mathbb{N}.$$

5.C.9. The answer is ± 1.

5.C.10. The result is

$$\limsup_{n\to\infty} a_n = 1, \qquad \liminf_{n\to\infty} a_n = 0.$$

5.C.11. We have

$$\liminf_{n\to\infty} \left((-1)^n \left(1 + \frac{1}{n} \right)^n + \sin \frac{n\pi}{4} \right) = -\mathrm{e} - \frac{\sqrt{2}}{2}.$$

5.D.5. The examined function is continuous on the whole $\mathbb{R}$.

5.D.6. The function is continuous at the points $-\pi$, 0, π; only right-continuous at the point 2; only left-continuous at the point 3; and continuous from neither side at 1.

5.D.7. It is necessary to set $f(0) := 0$.

5.D.8. The function is continuous iff $p = 2$.

5.D.9. The correct answer is $a = 4$.

5.D.10. It holds that

$$\lim_{x\to 0+} \frac{\sin^8 x}{x^3} = \lim_{x\to-\infty} \frac{\sin^8 x}{x^3} = 0.$$

5.D.13. The only solution is $x = -1$.

5.D.14. It has even two roots, because $P(-1) > 0 > P(0) < 0 < P(1)$ and due to 5.2.19 there is one root in $(-1, 0)$ and one in $(0, 1)$.

5.E.4. $f'(x) = 2x^{\ln x - 1} \cdot \ln x$.

5.E.5. $(\sin x)^{1+\cos x} \left(\cot^2 x - \ln (\sin x) \right)$.

5.E.7. $\frac{\pi}{6} - \frac{2}{\sqrt{3}} 0.003$.

5.E.8. $a \approx \frac{\pi}{4} + 0.01$; $b \approx 4.125$.

5.E.9. a) $\frac{1}{2} - \frac{\sqrt{3}\pi}{360}$; b) $\frac{\sqrt{2}}{2} + \frac{\sqrt{2}\pi}{360}$.

5.E.13. (a) $12\,\mathrm{ft/s}$; (b) $-59, 5\,\mathrm{ft^2/s}$; (c) $-1\,\mathrm{rad/s}$.

5.E.14. The slope of the tangent line of the polynomial P is given by the derivative of the polynomial. Consider $P(0)$ and $P(2)$. Yes, there is.

5.E.15. $y = \frac{x_0}{p} x - \frac{x_0^2}{2p}$.

5.E.16. $y = 2x$.

5.E.17. $y = \sqrt[3]{4}(x+1)$; $y = -\frac{\sqrt[3]{2}}{2}(x+1)$.

5.E.18. $y - \frac{\ln 5}{2} = \left(\frac{13}{10} - \frac{\ln 5}{4} \right)(x-1)$; $y - \frac{\ln 5}{2} = \frac{20}{5\ln 5 - 26}(x-1)$.

5.E.19. $\left[\frac{1}{2}, 2\frac{1}{4} \right]$.

5.E.20. $t: y = \frac{x}{6} + \frac{8}{3}$; $n: y = -6x + 15$; $\left[\frac{3}{2}, \sqrt{\left(\frac{3}{2} \right)^2 - 3\frac{3}{2} + 11} \right]$.

5.E.21. $\pi/4$.

5.E.22. $y = 2 - x$; $y = x$.

5.E.23. The inequalities follow, for instance, from the mean value theorem (attributed to Lagrange) applied to the function $y = \ln(1+t)$, $t \in [0, x]$.

5.I.6. $r = +\infty$.

5.I.7. 1.

5.I.8. 3.

5.I.9. $[-1, 1]$.

5.I.10. $x \in \left[2 - \frac{1}{3}, 2 + \frac{1}{3}\right]$.

5.I.11. It is.

5.I.12.

(a) True.
(b) False.
(c) False.
(d) True.

5.I.13. $1 - \frac{\pi^2}{10^2 \cdot 2} + \frac{\pi^4}{10^4 \cdot 4!}$.

5.I.14. The error lies in the interval $(0, 1/200)$.

5.I.15. $\sum_{n=0}^{\infty} \frac{e}{n!} (x-1)^n$; $\sum_{n=0}^{\infty} \frac{\ln^n 2}{n!} x^n$.

5.I.16. $f(x) = x$, $x \in \mathbb{R}$; it is.

5.I.17. It does not.

5.I.18. (a) $1 - \frac{\pi^2}{18^2 \cdot 2!} + \frac{\pi^4}{18^4 \cdot 4!}$; (b) $\frac{1}{2} - \frac{1}{5 \cdot 2^5}$.

5.I.19. $\sum_{n=0}^{\infty} \frac{1}{(2n+1)\, n!} x^{2n+1}$.

5.I.20. $a > 1$.

5.I.21. $[-\sqrt[3]{2}, \sqrt[3]{2})$.

5.I.22. For $x \in [-1, 1]$.

5.I.23. $x > 2$.

5.I.24. The series is absolutely convergent.

5.I.25. $\ln (3/2)$.

5.I.26. $\frac{x(1-x)}{(1+x)^3}$.

5.I.27. (a) $\frac{1}{2} \ln \frac{1+x}{1-x}$; (b) $\frac{1+x}{(1-x)^3}$.

5.I.28. $2/9$.

5.I.29. $x\, e^{\frac{x^3}{2}}$.

5.J.1. $x^4 + 2x^3 - x^2 + x - 2$.

5.J.2. $x^4 + 2x^3 - 2x^2 + x + 2$.

5.J.3. $x^4 + 3x^3 - 3x^2 - x - 1$.

5.J.4. For every $\varepsilon > 0$, it suffices to assign to the ε-neighborhood of the point -2 the δ-neighborhood of the point 0 given by

$$\varepsilon \mapsto \delta, \qquad \delta = \varepsilon,$$

and without loss of generality, we can assume that $\varepsilon \leq 1$. Since if $\varepsilon > 1$, we can set $\delta = 1$.

5.J.5. Existence of the limit and the equality

$$\lim_{x \to -1} \frac{(1+x)^2 - 3}{2} = -\frac{3}{2}$$

follows from the choice $\delta := \varepsilon$ for $\varepsilon \in (0, 1)$.

5.J.6. Since $-(x-2)^4 < x$ for $x < 0$, we get $3(x-2)^4/2 > -x$ for $x < 0$.

5.J.7. As

$$\lim_{x \to 0+} \arctan \frac{1}{x} = \frac{\pi}{2}, \qquad \lim_{x \to 0-} \arctan \frac{1}{x} = -\frac{\pi}{2},$$

the considered limit does not exist.

5.J.8. The former limit equals $+\infty$, the latter does not exist.

5.J.9. The limit can be determined in many ways. For instance:

$$\lim_{x\to0}\frac{\tan x-\sin x}{\sin^3 x}=\lim_{x\to0}\left(\frac{\tan x-\sin x}{\sin^3 x}\cdot\frac{\cot x}{\cot x}\right)$$
$$=\lim_{x\to0}\frac{1-\cos x}{\cos x\cdot\sin^2 x}=\lim_{x\to0}\frac{1-\cos x}{\cos x\,(1-\cos^2 x)}$$
$$=\lim_{x\to0}\frac{1}{\cos x\,(1+\cos x)}=\frac{1}{2}.$$

5.J.10. We have

$$\lim_{x\to\pi/6}\frac{2\sin^3 x+7\sin^2 x+2\sin x-3}{2\sin^3 x+3\sin^2 x-8\sin x+3}=\lim_{x\to\pi/6}\frac{\sin x+1}{\sin x-1}=-3.$$

5.J.11. We have

$$\lim_{x\to1}\frac{x^m-1}{x^n-1}=\frac{m}{n}.$$

5.J.12. After multiplying by the fraction

$$\frac{\sqrt{x^2+x}+x}{\sqrt{x^2+x}+x},$$

it follows that

$$\lim_{x\to+\infty}\left(\sqrt{x^2+x}-x\right)=\frac{1}{2}.$$

5.J.13. We have

$$\lim_{x\to+\infty}\left(x\sqrt{1+x^2}-x^2\right)=\frac{1}{2}.$$

5.J.14. We have

$$\lim_{x\to0}\frac{\sqrt{2}-\sqrt{1+\cos x}}{\sin^2 x}=\frac{\sqrt{2}}{8}.$$

5.J.15. By extending the given fraction, we obtain

$$\lim_{x\to0}\frac{\sin(4x)}{\sqrt{x+1}-1}=8.$$

5.J.16. We have

$$\lim_{x\to0-}\frac{\sqrt{1+\tan x}-\sqrt{1-\tan x}}{\sin x}=1.$$

5.J.17.

$$\lim_{x\to-\infty}\frac{2^x+\sqrt{1+x^2-x^9}-7x^5+44x^2}{3^x+\sqrt[5]{6x^6+x^2-18x^5-592x^4}}=\frac{7}{18}.$$

5.J.18. The statement is false. For example, consider

$$f(x):=\frac{1}{x},\quad x\in(-\infty,0);\qquad g(x):=x,\quad x\in\mathbb{R}.$$

5.J.19.

$$\lim_{n\to\infty}\left(\frac{n}{n+5}\right)^{2n-1}=e^{-10}.$$

5.J.20. $-\frac{1}{6}$.

5.J.21. $f'(x)<0$, $x>e$.

5.J.22. The function has a local maximum at the point $x_1=e^{-2}$. It has a local minimum at the point $x_2=1$.

5.J.23. No: if $a=\sqrt{2}/2$, there is only a local extremum at the point.

5.J.24. $2=e^{\frac{1}{e}}-\ln\frac{1}{e}$.

5.J.25. $\frac{1}{\sqrt[3]{e}}$.

5.J.26. $4=p(-1)=p(2)$, $-16=p(-3)$.

5.J.27. (a) $v(0)=6\,\text{m/s}$; (b) $t=3\,\text{s}$, $s(3)=16\,\text{m}$; (c) $v(4)=-2\,\text{m/s}$, $a(4)=-2\,\text{m/s}^2$.

5.J.28. $f'(x_0) = \frac{1}{2\sqrt{x_0}}$.

5.J.29. It does not because the one-sided derivatives differ (specifically: $\pi/2$ from the right and $-\pi/2$ from the left).

5.J.30. It does.

5.J.31. It does not.

5.J.32. $f(x) := |x - 5| + |x - 9|$.

5.J.33. Let $f = g = 1$ at the rational numbers and $f = g = -1$ at the irrational numbers.

5.J.34. (a) $x^2 \sin x$; (b) $\cos(\sin x) \cdot \cos x$; (c) $\frac{3x^2+2}{x^3+2x} \cos\left(\ln\left(x^3 + 2x\right)\right)$; (d) $\frac{2(1-2x)}{(1-x+x^2)^2}$.

5.J.35. (a) $\frac{7}{8} x^{-\frac{1}{8}}$; (b) $\operatorname{cosec} x = \frac{1}{\sin x}$.

5.J.36. $\cos x \cdot \cos(\sin x) \cdot \cos(\sin(\sin x))$.

5.J.37. $f'(x) = \frac{1}{\sqrt{1+2x-x^2}} + 1$, $x \in \left(1 - \sqrt{2}, 1 + \sqrt{2}\right)$.

5.J.38. $\frac{\cos x}{3\sqrt[3]{\sin^2 x}}$.

5.J.39. $\frac{1+2x^2}{\sqrt{1+x^2}} + x^2 e^x$.

5.J.40. -8.

5.J.41. $\frac{2x^2}{1-x^6} \sqrt[3]{\frac{1+x^3}{1-x^3}}$.

5.J.42. $\ln^2\left(x + \sqrt{1 + x^2}\right)$, $x \in \mathbb{R}$.

5.J.43. $f'(x) = -\frac{1}{x}\left(\log_x e\right)^2$, $x > 0$, $x \neq 1$.

5.J.44. $[f(x)g(x)h(x)k(x)]' = f'(x)g(x)h(x)k(x) + f(x)g'(x)h(x)k(x) + f(x)g(x)h'(x)k(x) + f(x)g(x)h(x)k'(x)$.

5.J.45. $\frac{x^3 (x+1)^2 \sqrt[3]{x+2}}{(x+3)^2}\left(\frac{3}{x} + \frac{2}{x+1} + \frac{1}{3(x+2)} - \frac{2}{x+3}\right)$.

5.J.52. 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.J.53. $4\,\mathrm{m} \times 4\,\mathrm{m} \times 2\,\mathrm{m}$.

5.J.54. $28 = 24 + 4$.

5.J.55. $a = 1$.

5.J.56. $2\sqrt{5}\,r$.

5.J.57. It is the square with sides of length c).

5.J.58. $h = \frac{4}{3}R$, $r = \frac{2\sqrt{2}}{3}R$.

5.J.59. It is the equilateral triangle (with area $\sqrt{3}\,p^2/36$).

5.J.60. $[2, -1/2]$, $[-2, -1/2]$.

5.J.61. $h = 2r$.

5.J.62. The closest point is $[1, 1]$, the distance then $2\sqrt{2}$.

5.J.63. The closest point is $[-1, 1]$, distance $3\sqrt{2}$.

5.J.64. $t = 1,5s$, the distance between will be $\sqrt{5}$ units.

5.J.65. It will happen at the time $t = \frac{5}{13}$ s, the distance being $\frac{\sqrt{13}}{13}$ units.

5.J.66. $P = \pi r v + \pi r^2 \implies v = \frac{P - \pi r^2}{\pi r} \implies V = \frac{1}{3}r(P - \pi r^2)$. The extremum is at $r = \sqrt{\frac{P}{3\pi}}$, the substitution gives $V = \frac{2\pi}{3}$ cm^3.

5.J.67. At about $3\,414$ products per day.

5.J.68. Triple use of l'Hospital's rule gives

$$\lim_{x \to 0-} \frac{\sin x - x}{x^3} = -\frac{1}{6}.$$

5.J.69. $2/\pi$.

5.J.70.

$$\lim_{x \to \frac{\pi}{2}-} \left(\frac{\pi}{2} - x\right) \tan x = 1.$$

5.J.71.
$$\lim_{x \to +\infty} \left(\left(3^{\frac{1}{x}} - 2^{\frac{1}{x}} \right) x \right) = \ln \frac{3}{2}.$$

5.J.72. $1/2$.

5.J.73. We have
$$\lim_{x \to +\infty} \left(\cos \frac{2}{x} \right)^{x^2} = e^{-2}.$$

5.J.74. By applying l'Hospital's rule twice, one obtains
$$\lim_{x \to 0} (1 - \cos x)^{\sin x} = e^0 = 1.$$

5.J.75. In both cases, the result is e^{α}.

5.J.76. The limit is easily calculated by l'Hospital's rule, for instance.

5.J.89. $2a^2$; $4a \left(2 + \sqrt{2} \right)$.

5.J.90. $\pi/2$.

5.J.91. $x = \frac{\pi}{6} + k\pi$, $x = \frac{5\pi}{6} + k\pi$, $k \in \mathbb{Z}$.

5.J.92. 5.

5.J.93. $+\infty$.

5.J.94. $3/2$.

5.J.95. (a) 3; (b) $9/4$.

5.J.96. $1/2$.

5.J.97. (a) $3/4$; (b) $1/4$.

5.J.98. $-1/2$.

5.J.99. $11/18$.

5.J.100. $s/2$; $3s/2$ $(s = \ln 2)$.

5.J.101. It does.

5.J.102. It suffices to consider the necessary condition for convergence, namely $\lim_{n \to \infty} a_n = 0$.

5.J.103. $\alpha > 0$; $\beta \in \{-2, -1, 0, 1, 2\}$; $\gamma \in (-\infty, -1) \cup (1, +\infty)$.

5.J.104. It is absolutely convergent.

5.J.105. The limit is $1/2$.

5.J.106. $A \in [0, 1)$.

5.J.107. The value of the given series is finite – the series converges.

5.J.108. For example: $a_n = n/3$, $b_n = n/2$, $n \in \mathbb{N}$.

5.J.109. The former series converges absolutely; the latter converges conditionally.

5.J.110. It does.

5.J.111. $p \in \mathbb{R}$.

5.J.112. $\mathbb{R}$.

5.J.113. $(1, e]$.

5.J.114. $\left(-\infty, \frac{2}{3} \right) \cup \left(\frac{2}{3}, +\infty \right)$; $\left(-\infty, -\frac{1}{3} \right) \cup \left(-\frac{1}{3}, +\infty \right)$; $y = \frac{2x+1}{3x+1}$, $x \neq -\frac{1}{3}$.

5.J.115. (a) yes; (b) no; (c) no; (d) no; (e) yes; (f) yes; (g) yes; (h) no.

5.J.116. (a) no; (b) no; (c) yes; (d) yes; (e) no; (f) no; (g) no; (h) yes.

5.J.117. The functions (a), (e) are odd; the functions (c), (d) are even.

5.J.118. It is periodic. The prime period is (a) 2π; (b) $\pi/3$.

5.J.119. The functions f and g are even, so it suffices to consider the graphs of the functions $y = e^x$, $x \in [0, +\infty)$ and $y = \ln x$, $x \in (0, +\infty)$.

5.J.120. The given function is even, so to draw its graph, it suffices to know the graph of the function $y = 2^x$, $x \in (-\infty, 0]$.

5.J.121. $(\sinh x)' = \cosh x$; $(\cosh x)' = \sinh x$; $(\tanh x)' = \frac{1}{\cosh^2 x}$; $(\coth x)' = -\frac{1}{\sinh^2 x}$.

5.J.122. $\frac{1}{\sqrt{1+x^2}}$.

5.J.123. Consider the series with the expansions of the functions sinh and cosh into power series. The result is

$$\sinh(1) + 2\cosh(1).$$

Differential and integral calculus

we already have the menagerie, but what shall we do with it?
– we'll learn to control it...

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 and complete several missing steps in the argumentation so far. We also add useful techniques, how to deal with extra parameters in the functions, and we introduce some further integration concepts briefly.

A. Derivatives of higher orders

First we'll introduce a convention for denoting the derivatives of higher orders: we'll denote the second derivative of function f of one variable by f'' or $f^{(2)}$, derivatives of third or higher order only by $f^{(3)}, f^{(4)}, \ldots f^{(n)}$. For remembrance, we'll start with a slightly cunning problem using "only" first derivatives.

6.A.1. Determine the following derivatives:

i) $(x^2 \cdot \sin x)''$,

ii) $(x^x)''$,

iii) $\left(\frac{x}{\ln x}\right)^{(3)}$,

iv) $(x^n)^{(n)}$,

v) $(\sin x)^{(n)}$.

Solution. (a) $(x^2 \cdot \sin x)'' = (2x \sin x + x^2 \cos x)'$
$= 2 \sin x + 4x \cos x - x^2 \sin x$.
(b) $(x^x)'' = [(1 + \ln x)x^x]' = x^{x-1} + x^x(1 + \ln x)^2$.
(c) $\left(\frac{x}{\ln x}\right)^{(3)} = \frac{1}{x^2(\ln x)^2} - \frac{6}{x^2(\ln x)^4}$.
(d) $(x^n)^{(n)} = [(x^n)']^{(n-1)} = (nx^{n-1})^{(n-1)} = \cdots = n!$.
(e) $(\sin x)^{(n)} = \operatorname{Re}(i^n \sin x) + \operatorname{Im}(i^n \cos x)$. $\qquad \square$

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 $0, 1, \ldots, \infty$. Often we write only C^k, if the domain is known from the context. When $k = 0$, C^0 means continuous functions.

6.A.2. Differentiate the expression

$$\frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}.$$

of variable $x > 1$.

Solution. We'll solve this problem using the so called logarithmic differentiation. Let f be an arbitrary positive function. We know that

$$[\ln f(x)]' = \frac{f'(x)}{f(x)}, \quad \text{tj.} \quad f'(x) = f(x)\cdot[\ln f(x)]',$$

if the derivative $f'(x)$ exists. The usefulness of this formula is given by the fact that for some functions, it's easier to differentiate their logarithm then themselves. Such is the expression in our provlem. We'll obtain

$$\left(\frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}\right)'$$

$$= \frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}\cdot\left[\ln\frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}\right]'$$

$$= \frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}$$

$$\cdot\left[3\ln(x+2) + \frac{1}{4}\ln(x-1) - x\ln\mathrm{e} - 2\ln(x+132)\right]'$$

$$= \frac{\sqrt[4]{x-1}\cdot(x+2)^3}{\mathrm{e}^x(x+132)^2}\left[\frac{3}{x+2} + \frac{1}{4(x-1)} - 1 - \frac{2}{x+132}\right].$$

$\square$

6.A.3. Let $n \in \mathbb{N}$ be arbitrary. Find the n-th derivative of function

$$y = \ln\frac{1+x}{1-x}, \quad x \in (-1,1).$$

Solution. With respect of the equality

$$\ln\frac{1+x}{1-x} = \ln(1+x) - \ln(1-x), \quad x \in (-1,1),$$

we'll define an auxiliary function

$$f(x) := \ln(ax+1), \quad x \in (-1,1), a = \pm 1.$$

For $x \in (-1,1)$ we can easily (sequentially) compute

$$f'(x) = \frac{a}{ax+1},$$

$$f''(x) = \frac{-a^2}{(ax+1)^2},$$

$$f^{(3)}(x) = \frac{2a^3}{(ax+1)^3},$$

$$f^{(4)}(x) = \frac{-6a^4}{(ax+1)^4}.$$

Based on these results we can figure out that

(1)

$$f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!\,a^n}{(ax+1)^n}, \quad x \in (-1,1),\, n \in \mathbb{N}.$$

We'll verify the validity of this formula by mathematical induction. It holds for $n = 1,2,3,4$, so it suffices to show that its validity for $k \in \mathbb{N}$ implies its validity for $k+1$. Because the direct computation yields

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, then 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 combinatorical 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 a 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.
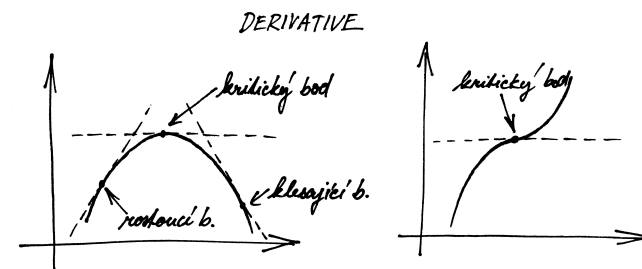
If x_0 is a critical point of function f, there are several possibilities for the behaviour of the function f in the neighbourhood of x_0. Consider the behaviour of the function $f(x) = x^n$ in the neighbourhood of zero for different

$$f^{(k+1)}(x) = \left(\frac{(-1)^{k-1}(k-1)!\,a^k}{(ax+1)^k} \right)' =$$
$$\frac{(-1)^{k-1}(k-1)!\,a^k\,(-k)\,a}{(ax+1)^{k+1}} = \frac{(-1)^k k!\,a^{k+1}}{(ax+1)^{k+1}},$$

(1) holds for all $n \in \mathbb{N}$. Then

$$\ln^{(n)}(1+x) = \frac{(-1)^{n-1}(n-1)!}{(x+1)^n}, \quad \ln^{(n)}(1-x) =$$
$$-\frac{(n-1)!}{(-x+1)^n}, \quad x \in (-1,1).$$

From here we obtain the result

$$\left(\ln \frac{1+x}{1-x} \right)^{(n)} = (n-1)! \left(\frac{1}{(1-x)^n} - \frac{(-1)^n}{(1+x)^n} \right)$$

pro $x \in (-1,1)$ a $n \in \mathbb{N}$. □

6.A.4. Determine the second derivative of function $y = \operatorname{tg} x$ on its whole domain, i.e. for $\cos x \neq 0$. ◯

6.A.5. Determine the fifth and the sixth derivative of the polynomial $p(x) = \left(3x^2 + 2x + 1\right) \cdot (2x - 6) \cdot \left(2x^2 - 5x + 9\right)$, $x \in \mathbb{R}$. ◯

6.A.6. With shorcuts determine the 12th derivative of function
$y = \mathrm{e}^{2x} + \cos x + x^{10} - 5x^7 + 6x^3 - 7x + 3$, $x \in \mathbb{R}$. ◯

6.A.7. Write the 26th derivative of function
$f(x) = \sin x + x^{23} - x^{18} + 5x^{11} - 3x^8 + \mathrm{e}^{2x}$, $x \in \mathbb{R}$. ◯

6.A.8. Let f be a given function and let z be a point such that

$$f(z) = 0, \quad f'(z) = 0, \quad f''(z) = 0, \quad f^{(3)}(z) = 1.$$

Which of the following statements:

(a) the tangent line to the graph of function f at point $[z, f(z)]$ is the x axis;

(b) the function f is not a polynomial of degree two;

(c) the function f is increasing at point z;

(d) the function f does not have a strict local minimum at point z ;

(e) the point z is an inflective point of function f

are necessarily true? ◯

6.A.9. Let a movement of a solid (a trajectory of a mass point) be described by function

$$s(t) = -(t-3)^2 + 16, \quad t \in [0,7]$$

in units m, s. Determine

(a) the initial (i.e. at time $t = 0$ s) velocity of the solid;

(b) the time and location at which the solid has zero velocity;

(c) the velocity and the acceleration of the solid at time $t = 4$ s.

Recall that velocity is the derivative of trajectory and acceleration is the derivative of velocity. ◯

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 of 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 return to the qualitative study of the behaviour of functions later on.

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*[1].

Consider the power series centered at a,

$$S(x) = \sum_{n=0}^{\infty} a_n (x-a)^n.$$

Differentiate it repeatedly, to get the power series

$$S^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1) \ldots (n-k+1) a_n (x-a)^{n-k}.$$

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

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{k!} S^{(k)}(a)(x-a)^n.$$

(NB. A power series can be differentiated term after term, This is proved later.)

---

[1] 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"

Taylor expansions. We necessarily need the derivatives of higher orders to determine the Taylor expansion of a given function.

6.A.10. Determine the Taylor expansions T_x^k (of k-th order at point x) of the following functions:

 i) T_0^3 of function $\sin x$,

 ii) T_1^3 of function $\frac{e^x}{x}$.

Solution. (i) We'll compute the values of the first, second and third derivative of function $f = \sin$ at point 0: $f'(0) = \cos(0) = 1$, $f^{(2)}(0) = -\sin(0) = 0$, $f^{(3)}(0) = -\cos(0) = -1$, also $f(0) = 0$. Thus the Taylor expansion of the third order of function$\sin(x)$ at point 0 is

$$T_0^3(\sin(x)) = x - \frac{1}{6}x^3.$$

(ii) Again $f(1) = e$,

$$f'(1) = \frac{e^x}{x} - \frac{e^x}{x^2}\bigg|_{x=1} = 0$$

$$f^{(2)}(1) = \frac{e^x}{x} - 2\frac{e^x}{x^2} + \frac{2e^x}{x^3}\bigg|_{x=1} = e$$

$$f^{(3)}(1) = \frac{e^x}{x} - 3\frac{e^x}{x^2} + \frac{6e^x}{x^3} - \frac{6e^x}{x^4}\bigg|_{x=1} = -2e$$

Thus we get the Taylor expansion of third order of function $\frac{e^x}{x}$ at point 1:

$$T_1^3(\frac{e^x}{x}) = e + \frac{e}{2}(x-1)^2 - \frac{e}{3}(x-1)^3 = e(-\frac{x^3}{3} + \frac{3x^2}{2} - 2x + \frac{5}{6}).$$

$\square$

6.A.11. Determine the Taylor polynomial T_0^6 of function $\sin$ and using theorem (6.1.3), estimate the error of the polynomial at point $\pi/4$.

Solution. Analogously to the previous example, we compute

$$T_0^6(\sin(x)) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5.$$

Using the theorem 6.1.3, we then estimate the size of the remainder (error) R. According to the theorem, there exists $c \in (0, \frac{\pi}{4})$ such that

$$R(\pi/4) = \left|\frac{-\cos(c)\pi^7}{7!4^7}\right| < \frac{1}{7!} \doteq 0,0002.$$

$\square$

6.A.12. Find the Taylor polynomial of third order of function

$$y = \operatorname{arctg} x, \quad x \in \mathbb{R}$$

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

$$T_{k,a}f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 +$$
$$\frac{1}{6}f^{(3)}(a)(x-a)^3 + \cdots + \frac{1}{k!}f^{(k)}(a)(x-a)^k.$$

The mean value theorem is used to show how good the approximation to f it 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*

$$f(x) = f(a) + f'(a)(x-a) + \dots$$
$$+ \frac{1}{(k-1)!}f^{(k-1)}(a)(x-a)^{k-1} + \frac{1}{k!}f^{(k)}(c)(x-a)^k$$
$$= T_{k-1,a}f(x) + \frac{1}{k!}f^{(k)}(c)(x-a)^k.$$

PROOF. For fixed x define the remainder R, by

$$f(x) = T_{k-1,a}f(x) + R.$$

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

$$F'(\xi) = f'(\xi) +$$
$$\sum_{j=1}^{k-1}\left(\frac{1}{j!}f^{(j+1)}(\xi)(x-\xi)^j - \frac{1}{(j-1)!}f^{(j)}(\xi)(x-\xi)^{j-1}\right)$$
$$- \frac{1}{(k-1)!}r(x-\xi)^{k-1}$$
$$= \frac{1}{(k-1)!}f^{(k)}(\xi)(x-\xi)^{k-1} - \frac{1}{(k-1)!}r(x-\xi)^{k-1}$$
$$= \frac{1}{(k-1)!}(x-\xi)^{k-1}(f^{(k)}(\xi) - r),$$

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

at point $x_0 = 1$.

6.A.13. Determine the Taylor expansion of third order at point $x_0 = 0$ of function

(a) $y = \frac{1}{\cos x}$;

(b) $y = e^{-\frac{x^2}{2}}$;

(c) $y = \sin(\sin x)$;

(d) $y = \operatorname{tg} x$;

(e) $y = e^x \sin x$

defined in a certain neighbourhood of point x_0.

6.A.14. Determine the Taylor expansion of fourth order of function $y = \ln x^2$, $x \in (0, 2)$ at point $x_0 = 1$.

6.A.15. 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.16. Write the Taylor polynomial of fourth degree of function $y = \sin x$, $x \in \mathbb{R}$ centered at the origin. Using this polynomial, approximately compute $\sin 1°$ and determine the limit

$$\lim_{x \to 0+} \frac{x \sin x - x^2}{x^4}.$$

6.A.17. Determine the Taylor polynomial centered at the origin of degree at least 8 of function $y = e^{2x}$, $x \in \mathbb{R}$.

6.A.18. Express the polynomial $x^3 - 2x + 5$ as a polynomial in variable $u = x - 1$.

We'll show some more interesting examples of using the differential calculus. First though, we'll mention the Jensen inequality, which disscusses convex and concave functions and which we'll use later.

6.A.19. Jensen inequality. For a strictly convex function f on interval I and for arbitrary points $x_1, \ldots, x_n \in I$ and real numbers $c_1, \ldots, c_n > 0$ sucht that $c_1 + \cdots + c_n = 1$, 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 occuring if and only if $x_1 = \cdots = x_n$.

Solution. Could be proven easily by induction: for $n = 2$ it is just the definition of the convex function, for the induction

exists a number c, $a < c < x$ such that $F'(c) = 0$. That is the desired relation. $\square$

A special case of the last theorem is the mean value theorem, as an approximation by Taylor series of degree zero. See (1).

6.1.4. Estimations for Taylor expansions. A simple case of a Taylor expansion is when f is a polynomial:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, \quad a_n \neq 0.$$

Because the $(n+1)$–th derivative f is identically zero, the Taylor polynomial of degree n has zero remainder, therefore for each $x_0 \in \mathbb{R}$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n.$$

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

$$|\sin x - (T_{k,0} \sin)(x)| \leq \frac{|x|^{k+1}}{(k+1)!}.$$

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 5.4.11 on the page 336.

As mentioned in the introduction of the discussion of Taylor expansion of functions, if we start with a power series $f(x)$ centered in a, then its partial sums coincide with Taylor polynomials $T_{k,a} f(x)$. 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.

step

$$f\left(\sum_{i=1}^{k+1} c_i\, x_i\right) = f\left(c_1 x_1 + (1-c_1)\sum_{i=2}^{k+1}\frac{c_i}{1-c_1}x_i\right)$$

$$\leq c_1 f(x_1) + (1-c_1)f\left(\sum_{i=2}^{k+1}\frac{c_i}{1-c_1}x_i\right)$$

$$\leq c_1 f(x_1) + (1-c_1)\left(\sum_{i=2}^{k+1}\frac{c_i}{1-c_1}f(x_i)\right)$$

$$= \sum_{i=1}^{k+1} c_i f(x_i),$$

where we used the inequality first for $n=2$ and then $n=k$. $\square$

Remark.

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.

6.A.20. 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

$$\sin\tfrac{\varphi_i}{2} = \tfrac{x_i}{r}, \quad \cos\tfrac{\varphi_i}{2} = \tfrac{h_i}{r}, \qquad i \in \{1,\dots,n\},$$

we have

$$S_i = x_i h_i = r^2 \sin\tfrac{\varphi_i}{2}\cos\tfrac{\varphi_i}{2} = \tfrac{1}{2}r^2\sin\varphi_i, i \in \{1,\dots,n\}.$$

This implies that the area of the hole n-gon is

$$S = \sum_{i=1}^{n} S_i = \tfrac{1}{2}r^2\sum_{i=1}^{n}\sin\varphi_i.$$

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

$$(1) \qquad \varphi_1 + \cdots + \varphi_n = \sum_{i=1}^{n}\varphi_i = 2\pi.$$

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

$$-\sin\left(\sum_{i=1}^{n}\tfrac{1}{n}\varphi_i\right) \leq$$

$$-\sum_{i=1}^{n}\tfrac{1}{n}\sin\varphi_i, \quad \text{tj.} \quad \sin\left(\sum_{i=1}^{n}\tfrac{1}{n}\varphi_i\right) \geq \sum_{i=1}^{n}\tfrac{1}{n}\sin\varphi_i.$$

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 $f(x)$.

Proof. The proof is identical with the 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

$$|f(x) - (T_{k,a}f)(x)| \leq \frac{M}{(k+1)!}|x|^{k+1}. \qquad \square$$

6.1.5. Analytic and smooth functions. If f is smooth at a, the formal power series can be written

$$S(x) = \sum_{n=0}^{\infty}\frac{1}{k!}f^{(k)}(a)(x-a)^n.$$

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.

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 are these numbers a_k.[2]

To show the essence of the problem, we introduce a function which has all its derivatives vanishing at zero, but is nonzero at every other point. We see later how useful this function is.

Consider the function defined by

$$f(x) = \mathrm{e}^{-1/x^2}.$$

It is a well defined smooth function at all points $x \neq 0$. Its limit at $x = 0$ exists, and $\lim_{x\to 0} f(x) = 0$. By defining $f(0) = 0$, f is a continuous function for all real x.
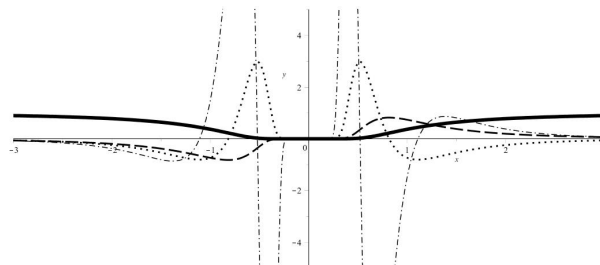
By a direct computation based on L'Hôpital's rule we compute the derivative of f (the first three ones are at the

---

[2]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.

Moreover, we know the equality occurs exactly for $\varphi_1 = \cdots = \varphi_n$. 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 $\varphi_1 = \cdots = \varphi_n$ (we chose $x_i = \varphi_i$). Hence the regular n-gon is the one with the maximum area, because it satisfies $\varphi_1 = \cdots = \varphi_n = 2\pi/n$. $\square$

6.A.21. Isoperimitric 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 occuring 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,

$$h = \cos \varphi = \cos \frac{\pi}{n}, \qquad \frac{x}{2} = \sin \varphi = \sin \frac{\pi}{n},$$

which yields

$$o_n = n \cdot x = 2n \sin \frac{\pi}{n}$$

and

$$S_n = n \cdot \frac{1}{2} h x = n \cos \frac{\pi}{n} \sin \frac{\pi}{n}.$$

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

picture, guess which line is which!). It suffices to consider only the right derivative, since the function is even.

$$f'(0) = \lim_{x \to 0^+} \frac{e^{-1/x^2} - 0}{x} = \lim_{x \to 0^+} \frac{x^{-1}}{e^{1/x^2}}$$
$$= \frac{1}{2} \lim_{x \to 0^+} \frac{x}{e^{1/x^2}} = 0.$$

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 $x \to 0^+$. This is a finite sum of limits of the expressions $x^{-j} e^{-1/x^2} = x^{-j}/e^{1/x^2}$. All these expressions are of type ∞/∞, 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.

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.

One of the most important functions in Mathematics and Physics is the *Gaussian function*

$$g(x) = f(x^{-1}) = e^{-x^2}.$$

We see why in Chapter 10 when dealing with probability and statistics.

Replace x with $-x^2$ in the power series for the exponential. It follows that g is an analytic function on the entire $\mathbb{R}$. Its derivatives are easily computed. 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. Both functions are depicted on the diagram. $f(x)$ is the solid line, while the Gaussian function is dashed.
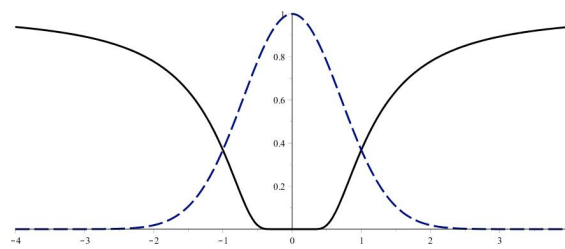
$$IQ = \tfrac{4\pi a^2}{(4a)^2} = \tfrac{\pi}{4} = \tfrac{\pi}{4}\cot g\,\tfrac{\pi}{4}.$$

Using the limit transition for $n \to \infty$ and the limit

$$\lim_{x\to 0}\tfrac{\sin x}{x} = 1,$$

we get the isoperimetric quotient for a circle:

$$IQ = \lim_{n\to\infty}\tfrac{\pi}{n}\cot g\,\tfrac{\pi}{n} = \lim_{n\to\infty}\frac{\cos\frac{\pi}{n}}{\frac{\sin\frac{\pi}{n}}{\frac{\pi}{n}}} = \frac{\cos 0}{1} = 1.$$

Of course, for a circle with radius r, we could have also directly computed

$$IQ = \tfrac{4\pi S}{o^2} = \tfrac{4\pi(\pi r^2)}{(2\pi r)^2} = 1.$$

For the boundary a of sector of a circle with radius r and central angle $\varphi \in (0, 2\pi)$, we have

$$IQ = \tfrac{4\pi S}{o^2} = \frac{4\pi\,\frac{\varphi r^2}{2}}{(2r+r\varphi)^2} = \frac{2\pi\varphi}{(2+\varphi)^2}.$$

Hence we're looking for a maximum of the function

$$f(\varphi) := \tfrac{2\pi\varphi}{(2+\varphi)^2}, \qquad \varphi \in (0, 2\pi).$$

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

$$f'(\varphi) > 0, \quad \varphi \in (0, 2), \qquad f'(\varphi) < 0, \quad \varphi \in (2, 2\pi).$$

Hence function f attains its maximal value for $\varphi_0 = 2$ and for a central angle $\varphi_0 = 2$ (radians), we get the largest

$$IQ = \tfrac{2\pi\varphi_0}{(2+\varphi_0)^2} = \tfrac{\pi}{4}.$$

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.∘ □

6.A.22. 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

$$\tfrac{1}{\lambda_i} := \tfrac{4\pi S_i}{o_i^2}, \qquad i \in \{1, \ldots, n\},$$

where S_i is the area and o_i the perimeter of the i-th figure. We'll also use the denotation

Notice how the function f touches the x axis due to the vanishing of all derivatives at zero.

6.1.6. Useful non-analytic smooth 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 function. 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:

$$g(x) = \begin{cases} 0 & \text{if } x \le 0 \\ e^{-1/x^2} & \text{if } x > 0. \end{cases}$$

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.

$$h(x) = \begin{cases} 0 & \text{if } |x| \ge a \\ e^{\frac{1}{x^2-a^2}+\frac{1}{a^2}} & \text{if } |x| < a. \end{cases}$$

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.
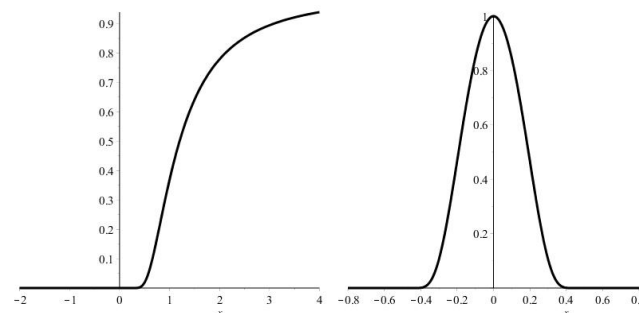
Finally we show how to get smooth analogies of the Heaviside functions. For two fixed real numbers $a < b$, define the function $f(x)$ exploiting the above function g as follows:

$$f(x) = \frac{g(x-a)}{g(x-a) + g(b-x)}.$$

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 $x \le a$ the numerator of the fraction is zero according to the definition of g. For $x \ge b$ 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 $\alpha = 0.4$.

$$\Lambda := \sum_{i=1}^{n} \lambda_i.$$

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 Λ 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

$$S_i = \frac{o_i^2}{4\pi\lambda_i}, \qquad i \in \{1, \dots, n\},$$

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 $c_1 + \cdots + c_n = 1$. Moreover we know that the equality occurs if and only if $x_1 = \cdots = x_n$. By choosing

$$c_i = \frac{\lambda_i}{\Lambda}, \quad x_i = \frac{o_i}{\lambda_i}, \qquad i \in \{1, \dots, n\},$$

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 occuring for

$$(1) \qquad x_1 = \cdots = x_n, \qquad \text{tj.} \qquad \frac{o_1}{\lambda_1} = \cdots = \frac{o_n}{\lambda_n}.$$

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.

$$o_i = \frac{\lambda_i}{\Lambda} l, \qquad i \in \{1, \dots, n\}.$$

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)
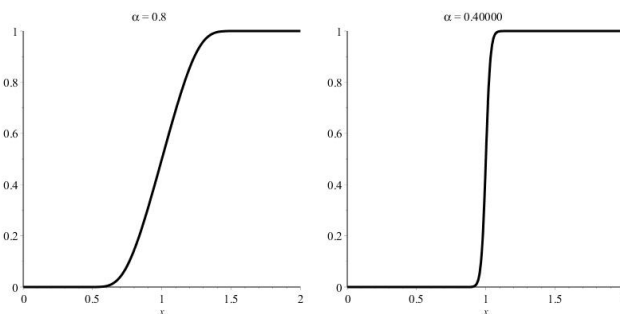
Finally, we can create a smooth analogue of the characteristic function of any interval $[c, d]$.

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 $(-\infty, c - \varepsilon)$ a $(d + \varepsilon, \infty)$. 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 $\varepsilon = 0.6$, $\varepsilon = 0.3$.
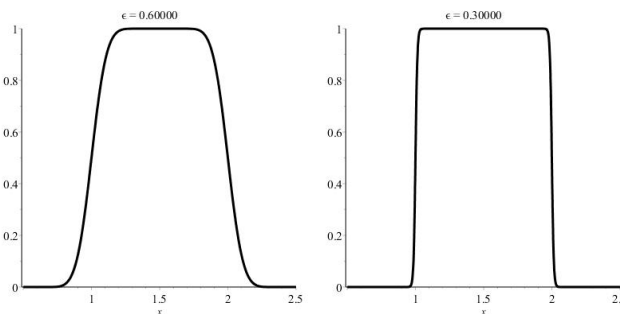
6.1.7. Local behaviour of functions. It is time to return to the behaviour of real functions of one real variable. We have seen that the sign of the first derivative of a differentiable function determines whether it is increasing or decreasing on some neighbourhood of the given point. If the derivative is zero, it does not of itself say much about the behaviour of the function.

We encountered the importance of the second derivative when describing critical points. 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

$$f'(a) = \cdots = f^{(k)}(a) = 0, \quad f^{(k+1)}(a) \neq 0.$$

$$\lambda_1 = \tfrac{4}{\pi}, \quad \lambda_2 = 1, \qquad \text{tj.} \qquad \Lambda = \lambda_1 + \lambda_2 = \tfrac{4+\pi}{\pi}.$$

Then the lengths of the respective parts are (in metres)

$$o_1 = \tfrac{\frac{4}{\pi}}{\frac{4+\pi}{\pi}} \cdot 1 = \tfrac{4}{4+\pi} \doteq 0,56, \quad o_2 = \tfrac{1}{\frac{4+\pi}{\pi}} \cdot 1 = \tfrac{\pi}{4+\pi} \doteq 0,44.$$

The area of a square with perimeter $0,56$ m (with a side of length $a = 0,14$ m) is $0,019\,6$ m^2 and the area of a circle with perimeter $0,44$ m (and radius $r \doteq 0,07$ m) is approximately $0,015\,4$ m^2. We can verify that (in m^2

$$\tfrac{l^2}{4\pi\Lambda} = \tfrac{1}{4(4+\pi)} \doteq 0,035 = 0,019\,6 + 0,015\,4.$$

$\square$

6.A.23. Expand the function

(a) $y = \ln \tfrac{1+x}{1-x}, \quad x \in (-1,1)$;

(b) $y = e^{x^2} + x^2 e^{-2x}, \quad x \in \mathbb{R}$

into a Taylor series centered at the origin..

Solution. If the function can be expressed as a sum of a power series (with a positive radius of convergence) on its domain of convergence, then this series is necessarily the Taylor series of the given function (its sum). This allows us to find the corresponding Taylor series easily.

Case (a). We know that

$$\ln(1+x) = \sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n} x^n, \quad x \in (-1,1),$$

i.e. for $x \in (-1,1)$ we have

$$\ln(1-x) = \sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n}(-x)^n = -\sum_{n=1}^{\infty} \tfrac{1}{n} x^n.$$

In total, for $x \in (-1,1)$ we have

$$\ln \tfrac{1+x}{1-x} = \ln(1+x) - \ln(1-x) = \sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}+1}{n} x^n$$

$$= \sum_{n=1}^{\infty} \tfrac{2}{2n-1} x^{2n-1}.$$

Case (b). Similarly, the well known identity

$$e^x = \sum_{n=0}^{\infty} \tfrac{1}{n!} x^n, \quad x \in \mathbb{R},$$

implies

$$e^{x^2} = \sum_{n=0}^{\infty} \tfrac{1}{n!} \left(x^2\right)^n = \sum_{n=0}^{\infty} \tfrac{1}{n!} x^{2n}, \quad x \in \mathbb{R},$$

and

$$x^2 e^{-2x} = x^2 \sum_{n=0}^{\infty} \tfrac{1}{n!} (-2x)^n = \sum_{n=0}^{\infty} \tfrac{(-2)^n}{n!} x^{n+2}, \quad x \in \mathbb{R}.$$

Hence

$$e^{x^2} + x^2 e^{-2x} = \sum_{n=0}^{\infty} \tfrac{x^{2n} + (-2)^n x^{n+2}}{n!}, \quad x \in \mathbb{R}.$$

$\square$

Suppose $f^{(k+1)}(a) > 0$. 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

$$f(x) = f(a) + \frac{1}{(k+1)!} f^{(k+1)}(c)(x-a)^{k+1}$$

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.

Similarly, if $f^{(k+1)}(a) < 0$, then it is a local maximum for odd k, and there is no extreme for even k.

6.1.8. 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,

$$f(x) \le f(a) + f'(a)(x-a).$$

Similarly f is *convex* at a, if its graph is above the tangent at the point a. That is,

$$f(x) \ge f(a) + f'(a)(x-a).$$

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 this point.
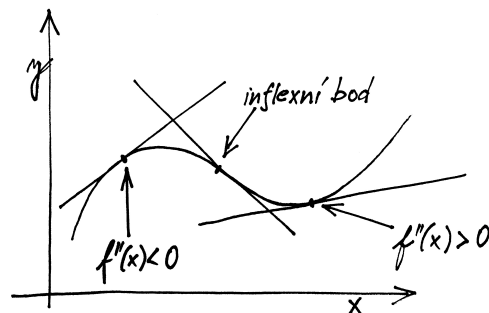
6.A.24. Determine the Taylor series centered at the origin of function

(a) $y = \frac{1}{(1+x)^2}$, $x \in (-1, 1)$;

(b) $y = \text{arctg}\, x$, $x \in (-1, 1)$.

Solution. Case (a). We'll use the formula

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n = \sum_{n=0}^{\infty} (-1)^n x^n, \quad x \in (-1, 1)$$

for the sum of a geometric series. By differentiating it, we obtain for $x \in (-1, 1)$

$$-\frac{1}{(1+x)^2} = \left(\sum_{n=0}^{\infty} (-1)^n x^n \right)' = \sum_{n=1}^{\infty} (-1)^n\, n\, x^{n-1},$$

with $\left(x^0 \right)' = 0$, thus the lower index is $n = 1$. We can see that

$$\frac{1}{(1+x)^2} = \sum_{n=1}^{\infty} (-1)^{n+1}\, n\, x^{n-1}, \quad x \in (-1, 1).$$

Case (b). We can express the derivative of function $y = \text{arctg}\, t$ for $t \in (-1, 1)$ as

$$(\text{arctg}\, t)' = \frac{1}{1+t^2} = \sum_{n=0}^{\infty} \left(-t^2 \right)^n = \sum_{n=0}^{\infty} (-1)^n t^{2n},$$

Because for $x \in (-1, 1)$ we have

$$\int_0^x (\text{arctg}\, t)'\, dt = \text{arctg}\, x - \text{arctg}\, 0 = \text{arctg}\, x$$

and

$$\int_0^x \left(\sum_{n=0}^{\infty} (-1)^n t^{2n} \right) dt = \sum_{n=0}^{\infty} \left((-1)^n \int_0^x t^{2n}\, dt \right)$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1},$$

we already have the result

$$\text{arctg}\, x = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}, \quad x \in (-1, 1).$$

$\square$

6.A.25. Find the Taylor series centered at $x_0 = 0$ of function

$$f(x) = \int_0^x u \cos u^2\, du, \quad x \in \mathbb{R}.$$

Solution. The equality

$$\cos t = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} t^{2n}, \quad t \in \mathbb{R}$$

implies

$$u \cos u^2 = u \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} \left(u^2 \right)^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} u^{4n+1}, \quad u \in \mathbb{R}$$

and then (for $x \in \mathbb{R}$)

$$f(x) = \int_0^x u \cos u^2\, du = \int_0^x \left(\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} u^{4n+1} \right) du$$

$$= \sum_{n=0}^{\infty} \left(\frac{(-1)^n}{(2n)!} \int_0^x u^{4n+1}\, du \right) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!\,(4n+2)} x^{4n+2}.$$

6.1.9. 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. 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.

6.1.10. Asymptotes of the graph of the function. We introduce one more useful utility for understanding/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 $x \to x_0$.

Thus, an asymptote at the improper point ∞ is a line $y = ax + b$, which satisfies

$$\lim_{x \to \infty} (f(x) - ax - b) = 0.$$

An *asymptote with a slope*. If such an asymptote exists, it satisfies

$$\lim_{x \to \infty} (f(x) - ax) = b$$

Consequently the limit

$$\lim_{x \to \infty} \frac{f(x)}{x} = a$$

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 rational functions have asymptotes at all zero points of the denominator which are zero points of the numerator as well.

□

6.A.26. Approximately compute $\cos\frac{\pi}{10}$ with an error lesser than 10^{-5}. ○

6.A.27. Without computing the derivatives, determine the Taylor polynomial of degree 4 centered at $x_0 = 0$ of function

$$f(x) = \cos x - 2\sin x - \ln(1 + x), \quad x \in (-1, 1).$$

Then decide if the graph of function f in neighbourhood of the point $[0, 1]$ is above or below the tangent line. ○

Now we'll state several "classical" problems, in which we'll determine the course of distinct functions. By determining the course we mean

(a) the domain (it's given) and the range;

(b) eventual parity and periodicity;

(c) discontinuities and their kind (including the according one-sided limits);

(d) points of intersections with the axes x, y;

(e) the intervals where the function is positive and where it's negative;

(f) the limits $\lim_{x\to-\infty} f(x)$, $\lim_{x\to+\infty} f(x)$;

(g) the first and the second derivatice;

(h) the critical and the so called stationary points, at which the first derivative is zero (eventually the points, at which the first or the second derivative don't exist)

(i) the intervals of monotonicity;

(j) strict and nonstrict local and absolute extremes;

(k) the intervals where the function is convex and where it's concave;

(l) the points of inflection;

(m) the horizontal and inclined asymptotes;

(n) values of the function f and its derivative f' at „significant" points;

(o) the graph.

6.A.28. Determine the range of function

$$f(x) = \frac{e^x - 1}{e^x + 1}, \quad x \in \mathbb{R}.$$

Solution. The line $y = 1$ is clearly an asymptote of function f at $+\infty$ and the line $y = -1$ is an asymptote at $-\infty$, because

$$\lim_{x\to\infty}\frac{e^x-1}{e^x+1} = \lim_{x\to\infty}\frac{e^x}{e^x} = 1, \quad \lim_{x\to-\infty}\frac{e^x-1}{e^x+1} = \frac{0-1}{0+1} = -1.$$

The inequality

$$f'(x) = \frac{2e^x}{(e^x+1)^2} > 0, \quad x \in \mathbb{R}$$

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 ± 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.
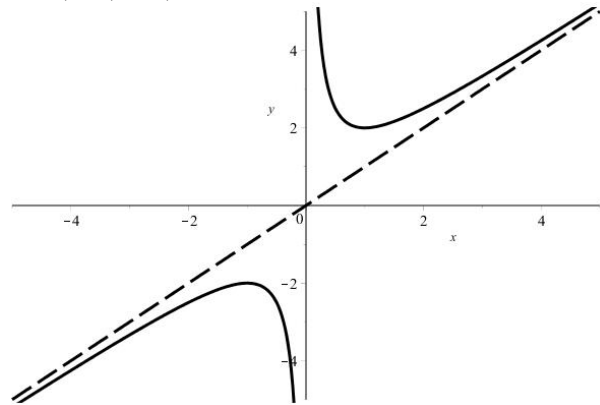
6.1.11. Differential of a function. In practical use of differential calculus, we often work with dependencies between several variables, say y and x. The choice of dependent and independent variable is not fixed. The explicit relation $y = f(x)$ with some function f is then only one of possible options. 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

$$df(x) = f'(x)dx,$$

where we interpret $df(x)$ as a linear map $\mathbb{R} \to \mathbb{R}$ defined on increments of x at the given point, $df(x)(v) = f'(x)\cdot v$, while $dx(x)(v) = v$.

We talk about the *differential of function* f if the following approximation property is true:

$$\lim_{v\to 0}\frac{f(x+v) - f(x) - df(x)(v)}{v} = 0.$$

Clearly, the differential is unique if it exists and, indeed, $df(x) = f'(x)dx$. Taylor theorem then implies that the differential df exists if the derivative f' is bounded. In particular, this happens at the point x if the first derivative $f'(x)$ exists and is continuous at x.

If the quantity x is expressed by another quantity t, e.g. $x = g(t)$ and, moreover, g has continuous first derivatives

then implies that f is continuous and increasing on $\mathbb{R}$. Hence the range is the interval $(-1, 1)$. □

6.A.29. Find all intervals on which the function $y = e^{-x^2}$, $x \in \mathbb{R}$ is concave. ◯

6.A.30. Consider function
$$y = \operatorname{arctg} \tfrac{x-1}{x}, \quad x \neq 0 \,(x \in \mathbb{R}).$$
Determine intervals on which this function is convex and concave and also all its asymptotes. ◯

6.A.31. Find all asymptotes of function

(a) $y = x\,e^x$;
(b) $y = \frac{(x+3)^3}{(x-2)^3}$

with maximal domain. ◯

6.A.32. Find the asymptotes of function
$$y = 2\operatorname{arctg}\left|\tfrac{x}{x^2-1}\right|, \quad x \neq \pm 1 \,(x \in \mathbb{R}).$$
◯

6.A.33. Consider function
$$y = \ln \tfrac{3e^{2x}+e^x+10}{e^x+1}$$
defined for all real x. Find its asymptotes. ◯

6.A.34. Determine the course of the function
$$f(x) = \sqrt[3]{|x|^3 + 1}.$$

Solution. The domain is the whole real axis, f has no discontinuities. For example it suffices to consider that the function $y = \sqrt[3]{x}$ is continuous at every point $x \in \mathbb{R}$ (unlike even roots defined only on the nonnegative axis). We can also immediately see that $f(x) \geq 1$ and $f(-x) = f(x)$ for all $x \in \mathbb{R}$, i.e. the function f is positive and even. Thus we can obtain the point $[0, 1]$ as the only intersections of the graph of f with the axes by substituting $x = 0$. The limit behavior of the function can be determined only at $\pm\infty$ (there are no discontinuities), where we can easily compute

(1)
$$\lim_{x \to \pm\infty} \sqrt[3]{|x|^3 + 1} = \lim_{x \to \pm\infty} \sqrt[3]{|x|^3} = \lim_{x \to \pm\infty} |x| = +\infty.$$

Now we'll step up to determing the course of a function by using its derivatives. For $x > 0$, we have
$$f(x) = \sqrt[3]{x^3 + 1} = \left(x^3 + 1\right)^{\frac{1}{3}},$$
hence

(2)
$$f'(x) = \frac{1}{3}\left(x^3 + 1\right)^{-\frac{2}{3}} 3x^2 = \frac{x^2}{\sqrt[3]{(x^3+1)^2}} > 0, \qquad x > 0.$$

again, the chain rule for differentiating the composite functions says that $f \circ g$ has the differential too and

$$df(t) = d(f \circ g)(t) = \frac{df}{dx}(x)\frac{dx}{dt}(t)dt.$$

Therefore df can be seen as a linear approximation of the given quantity dependent on the increments of the dependent variable, no matter how this dependence is given.

6.1.12. The curvature of the graph of a function. We shall conclude this section with two straightforward applications of differentials. First we discuss curves in the plane and space, starting with the graphs of functions. Then we provide a brief introduction to the numerical procedures for differentiation (jump to 6.1.16 if getting tired with the curvatures).

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 overview.

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.

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.

This implies that f is increasing on the interval $(0, +\infty)$. With respect to its continuity at the origin, it must be increasing on $[0, +\infty)$. Because it's an even function, we know that on the interval $(-\infty, 0]$ it must be decreasing. Thus it has only one local minimum at point $x_0 = 0$, which is also a (strict) global minimum. Because a nonconstant continuous function maps an interval to an interval, the range of f is exactly $[1, +\infty)$ (consider $f(x_0) = 1$ and (1)). Notice that thanks to the even parity of the function, we didn't have to compute the derivative f' on the negative half-axis, which can though be easily determined by substituting $|x|^3 = (-x)^3 = -x^3$, yielding

$$f'(x) = \tfrac{1}{3}\left(-x^3 + 1\right)^{-\frac{2}{3}}\left(-3x^2\right) = -\frac{x^2}{\sqrt[3]{(-x^3+1)^2}} < 0, \qquad x < 0.$$

When computing $f'(0)$, we can proceed according to the definition or we can use the limits

$$\lim_{x \to 0+} \frac{x^2}{\sqrt[3]{(x^3+1)^2}} = 0 = \lim_{x \to 0-} -\frac{x^2}{\sqrt[3]{(-x^3+1)^2}}$$

determine the one-sided derivatives and then $f'(0) = 0$. In fact, we didn't even have to compute the first derivative on the positive half-axis either. To obtain that f is increasing on $(0, +\infty)$, we only needed to realize that both functions $y = \sqrt[3]{x}$ and $y = x^3 + 1$ are increasing on $\mathbb{R}$ and a composition of increasing functions is again an increasing function.

For $x > 0$, we can easily compute the second derivative using (2)

$$f''(x) = \frac{2x\sqrt[3]{(x^3+1)^2} - \frac{2}{3}x^2\sqrt[3]{(x^3+1)^{-1}}(3x^2)}{\sqrt[3]{(x^3+1)^4}},$$

i.e. after a simplification we have

(3) $$f''(x) = \frac{2x}{\sqrt[3]{(x^3+1)^5}} > 0, \qquad x > 0.$$

Similarly we can compute $f''(x)$

$$= -\frac{2x\sqrt[3]{(-x^3+1)^2} - \frac{2}{3}x^2\sqrt[3]{(-x^3+1)^{-1}}\left(-3x^2\right)}{\sqrt[3]{(-x^3+1)^4}}$$

$$= -\frac{2x}{\sqrt[3]{(-x^3+1)^5}} > 0,$$

for $x > 0$ and then $f''(0) = 0$. Next, we can use a limit transition:

$$\lim_{x \to 0+} \frac{2x}{\sqrt[3]{(x^3+1)^5}} = 0 = \lim_{x \to 0-} -\frac{2x}{\sqrt[3]{(-x^3+1)^5}}.$$

According to the inequality (3), f is strictly convex on the interval $(0, +\infty)$. Also f must be strictly convex on $(-\infty, 0)$.
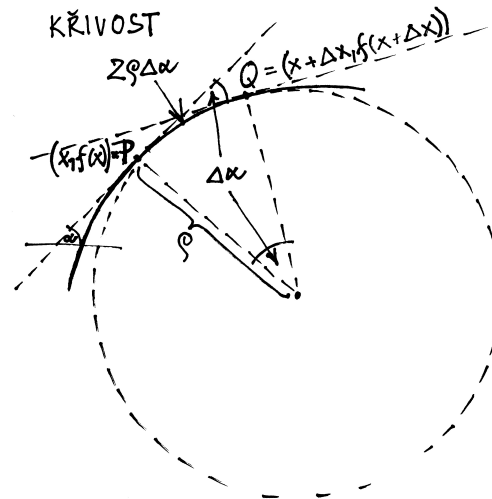
It can be seen from the figure that if the angle between the tangent at the fixed point P and the x axis is α and the angle between the tangent at the chosen point Q and the x axis is $\alpha + \Delta\alpha$, then the angle of the latter perpendicular lines is $\Delta\alpha$ as well.

Denote the radius of the circle by ρ. 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 Δs of the curve between P and Q, that is, the graph of the function $f(x)$. At the same time the circle approaches some circle C_P. Thus we arrive at the basic relation for the expected radius ρ of the circle C_P in terms of the linear approximations of the quantities:

$$\rho = \lim_{\Delta\alpha \to 0} \frac{\Delta s}{\Delta\alpha} = \frac{ds}{d\alpha}.$$

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 ρ.

For computing the radius ρ in terms of f we need to express the length of the arc s by the change of the angle α and express the derivative of this function by the derivative of f.

Notice that for an increasing angle α 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 ρ 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, $\mathrm{tg}\,\alpha = \frac{df}{dx}$. 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

$$\frac{1}{(\cos\alpha)^2} = 1 + (\mathrm{tg}\,\alpha)^2 = 1 + (f')^2$$

To obtain this conclusion though, we again didn't have to compute the second derivative for $x < 0$, it sufficed to use the even parity of the function. In total, we obtained that f is convex on its whole domain (it doesn't have any inflection points).

To be able to plot the graph of the function, we still need to find the asymptotes (we leave the computation of values of the function at certain points to the reader). Since f is continuous on $\mathbb{R}$, it can't have any horizontal asymptotes. A line $y = ax + b$ is an inclined asymptote for $x \to \infty$ if and only if both (proper) limits

$$\lim_{x \to \infty} \frac{f(x)}{x} = a, \qquad \lim_{x \to \infty} (f(x) - ax) = b.$$

exist. Analogous statement holds for $x \to -\infty$. Hence the limits

$$\lim_{x \to \infty} \frac{f(x)}{x} = \lim_{x \to \infty} \frac{\sqrt[3]{x^3+1}}{x} = \lim_{x \to \infty} \frac{\sqrt[3]{x^3}}{x} = 1,$$

$$\lim_{x \to \infty} (f(x) - 1 \cdot x) = \lim_{x \to \infty} \left(\sqrt[3]{x^3 + 1} - x \right) =$$

$$\lim_{x \to \infty} \left(\left[\sqrt[3]{x^3 + 1} - x \right] \frac{\sqrt[3]{(x^3+1)^2}+x\sqrt[3]{x^3+1}+x^2}{\sqrt[3]{(x^3+1)^2}+x\sqrt[3]{x^3+1}+x^2} \right) =$$

$$\lim_{x \to \infty} \frac{x^3+1-x^3}{\sqrt[3]{(x^3+1)^2}+x\sqrt[3]{x^3+1}+x^2} = \lim_{x \to \infty} \frac{1}{3x^2} = 0$$

imply that the line $y = x$ is an asymptote at $+\infty$. If we again consider the fact that f is even, we'll immediately obtain the line $y = -x$ as an asymptote at $-\infty$. $\quad\square$

6.A.35. Determine the course of the function

$$f(x) = \frac{\cos x}{\cos 2x}.$$

Solution. The domain consists of exactly those $x \in \mathbb{R}$, for which $\cos 2x \neq 0$. The equality $\cos 2x = 0$ is satisfied exactly for

$$2x = \tfrac{\pi}{2} + k\pi, \, k \in \mathbb{Z}, \quad \text{tj.} \quad x = \tfrac{\pi}{4} + \tfrac{k\pi}{2}, \, k \in \mathbb{Z}.$$

Hence the domain is

$$\mathbb{R} \smallsetminus \left\{ \tfrac{\pi}{4} + \tfrac{k\pi}{2}; \, k \in \mathbb{Z} \right\}.$$

Clearly we have

$$f(-x) = \frac{\cos(-x)}{\cos(-2x)} = \frac{\cos x}{\cos 2x} = f(x)$$

for all x in the domain, thus f (with its domain symmetric with respect to the origin) is an even function, which was implied by the even parity of the function $y = \cos x$. Moreover, because cosine is periodic with a period of 2π (i.e. $y = \cos 2x$ has a period of π), it suffices to consider the function f for

$$x \in \mathcal{D} := [0, \pi] \smallsetminus \left\{ \tfrac{\pi}{4} + \tfrac{k\pi}{2}; \, k \in \mathbb{Z} \right\} =$$
$$\left[0, \tfrac{\pi}{4}\right) \cup \left(\tfrac{\pi}{4}, \tfrac{3\pi}{4}\right) \cup \left(\tfrac{3\pi}{4}, \pi\right],$$

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,

$$\rho = \frac{ds}{d\alpha} = \frac{ds}{dx}\frac{dx}{d\alpha} = \frac{(1 + (f')^2)^{3/2}}{f''}.$$

The result explains the relation between the curvature and the second derivative. The numerator of the fraction is always positive. It equals the third power of the length of the tangent vector of the given 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.

If the second derivative is zero, the curvature $1/\rho$ is also zero. If f'' is large, then the radius ρ is small, thus the curvature is large as well.

The circle, by which curvature is defined is called the *osculating circle*.

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.

6.1.13. 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 $f : \mathbb{R} \to V$ need to have values in a vector space V. We call them *vector functions of one real variable* or more briefly *vector functions*.

To end this section, we digress to consider functions with values in the plane or in space. Thus, $f : \mathbb{R} \to \mathbb{R}^2$ and $f : \mathbb{R} \to \mathbb{R}^3$. 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

$$r(t) = x(t)e_1 + y(t)e_2, \quad r(t) = x(t)e_1 + y(t)e_2 + z(t)e_3.$$

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.

since the course of the function on its whole domain can be derived using its even parity and periodicity with a period of 2π.

Hence we'll only be concerned with the discontinuities $x_1 = \pi/4$ and $x_2 = 3\pi/4$. We'll determine the corresponding one-sided limits

$$\lim_{x \to \frac{\pi}{4}-} \frac{\cos x}{\cos 2x} = +\infty, \qquad \lim_{x \to \frac{\pi}{4}+} \frac{\cos x}{\cos 2x} = -\infty,$$

$$\lim_{x \to \frac{3\pi}{4}-} \frac{\cos x}{\cos 2x} = +\infty, \qquad \lim_{x \to \frac{3\pi}{4}+} \frac{\cos x}{\cos 2x} = -\infty.$$

If we have a respect to the continuity of f on the interval $(\pi/4, 3\pi/4)$, we can see that f attains all real values on this interval. Hence the range of f is the whole $\mathbb{R}$. We also found out that the discontinuities are of the second kind, where at least one of the one-sided limits is improper (or doesn't exist). By that, we simultaneously proved that the lines $x = \pi/4$ and $x = 3\pi/4$ are horizontal asymptotes. If we'd want to formulate the previous results without a restriction to the $[0, \pi]$, we can say that at all points

$$\hat{x}_k = \tfrac{\pi}{4} + \tfrac{k\pi}{2}, \qquad k \in \mathbb{Z}$$

f has a discontinuity of the second kind and every line

$$x = \tfrac{\pi}{4} + \tfrac{k\pi}{2}, \qquad k \in \mathbb{Z}$$

is a horizontal asymptote. Also the periodicity of f implies that no other asymptotes exist. In particular, it cannot have any inclined asymptotes, nor can the (improper) limits $\lim_{x \to +\infty} f(x)$, $\lim_{x \to -\infty} f(x)$ exist. Now we'll find the points of intersection with the axes. The point of intersection $[0, 1]$ with the y axis can be cound by computing $f(0) = 1$. When looking for the points of intersection with the x axis, we consider the equation $\cos x = 0$, $x \in \mathcal{D}$ with the only solution being $x = \pi/2$. Then we can easily obtain the intervals $[0, \pi/4), (\pi/2, 3\pi/4)$, on which f is positive, and the intervals $(\pi/4, \pi/2), (3\pi/4, \pi]$, where it's negative.

Now we'll step up to computing the derivative

$$
\begin{aligned}
f'(x) &= \frac{-\sin x \cos 2x - 2\cos x \left(-\sin 2x\right)}{\cos^2 2x} \\
&= \frac{-\sin x \left(\cos^2 x - \sin^2 x\right) + 2\cos x \left(2\sin x \cos x\right)}{\cos^2 2x} \\
&= \frac{\sin^3 x + 3\cos^2 x \sin x}{\cos^2 2x} \\
&= \frac{\left(\sin^2 x + \cos^2 x + 2\cos^2 x\right)\sin x}{\cos^2 2x} \\
&= \frac{\left(2\cos^2 x + 1\right)\sin x}{\cos^2 2x}, \ x \in \mathcal{D}.
\end{aligned}
$$

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 $r(t) = a\cos te_1 + a\sin te_2 + be_3$ 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.

6.1.14. 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 2.3.22 a 4.1.22.

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 $r(t) : \mathbb{R} \to \mathbb{R}^n$ is a differentiable vector and $\Psi : \mathbb{R}^n \to \mathbb{R}^m$ 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) Let there be differentiable vectors $r_1, \ldots, r_k : \mathbb{R} \to \mathbb{R}^n$ and a k–linear form $\Phi : \mathbb{R}^n \times \ldots \times \mathbb{R}^n$ on the space $\mathbb{R}^n$. Then the derivative of the composed map

$$\varphi(t) = \Phi(r_1(t), \ldots, r_k(t))$$

satisfies the (generalized) Leibniz rule

$$\frac{d\varphi}{dt} = \Phi\big(\frac{dr_1}{dt}, r_2, \ldots, r_k\big) + \cdots + \Phi\big(r_1, \ldots, r_{k-1}, \frac{dr_k}{dt}\big).$$

(3) The previous statement remains valid even if Φ 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

$$\Psi \circ r(t) = \left(\sum_{i=1}^{n} a_{1i}r_i(t), \ldots, \sum_{i=1}^{n} a_{mi}r_i(t)\right).$$

The points at which $f'(x) = 0$ are clearly the solutions of the equation $\sin x = 0$, $x \in \mathcal{D}$, i.e. the derivative is zero at points $x_3 = 0$, $x_4 = \pi$. The inequalities

$$2\cos^2 x + 1 \geq \cos^2 2x > 0, \quad \sin x > 0, \qquad x \in \mathcal{D} \cap (0, \pi)$$

imply that f is increasing at every inner point of the set $\mathcal{D}$, thus f is increasing on every subinterval of $\mathcal{D}$. The even parity of f then implies that it's decreasing at every point $x \in (-\pi, 0)$, $x \neq -3\pi/4$, $x \neq -\pi/4$. Hence the function has strict local extremes exactly at the points

$$\tilde{x}_k = k\pi, \qquad k \in \mathbb{Z}.$$

With respect to periodicity of f, we uniquely describe these extremes by stating that for $x_3 = \tilde{x}_0 = 0$, we get a local minimum (recall the value of the function $f(0) = 1$) and for $x_4 = \tilde{x}_1 = \pi$, a local maximum with the value $f(\pi) = -1$.

Let's compute the second derivative

$$f''(x) = \frac{[4\cos x(-\sin x)\sin x + (2\cos^2 x + 1)\cos x]\cos^2 2x}{\cos^4 2x}$$

$$-\frac{4\cos 2x(-\sin 2x)(2\cos^2 x + 1)\sin x}{\cos^4 2x} = \dots$$

$$= \frac{[10\sin^2 x \cos^2 x + 2\cos^4 x + \cos^2 x + 4\sin^4 x + 7\sin^2 x]\cos x}{\cos^3 2x},$$

$x \in \mathcal{D}$. Note that after a few simplifications, we can also express

$$f''(x) = \frac{(3 + 4\cos^2 x \sin^2 x + 8\sin^2 x)\cos x}{\cos^3 2x}, \qquad x \in \mathcal{D}$$

or

$$f''(x) = \frac{(11 - 4\cos^4 x - 4\cos^2 x)\cos x}{\cos^3 2x}, \qquad x \in \mathcal{D}.$$

Since

$$10\sin^2 x \cos^2 x + 2\cos^4 x + \cos^2 x + 4\sin^4 x + 7\sin^2 x >$$
$$0, \quad x \in \mathbb{R},$$

or

$$3 + 4\cos^2 x \sin^2 x + 8\sin^2 x = 11 - 4\cos^4 x - 4\cos^2 x \geq$$
$$3, \quad x \in \mathbb{R}$$

respectively, we have $f''(x) = 0$ for certain $x \in \mathcal{D}$ if and only if $\cos x = 0$. But that's satisfied only by $x_5 = \pi/2 \in \mathcal{D}$. It's clear that f'' changes its sign at this point, i.e. it's a point of inflection. No other points of inflection exist (the second derivative f'' is continuous on $\mathcal{D}$). Other changes of the sign of f'' occur at zero points of the denominator, which we have already determined as discontinuities $x_1 = \pi/4$ and $x_2 = 3\pi/4$. Hence the sign changes exactly at points x_1, x_2, x_5, thus the inequality

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 Ψ 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, …, 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 $B_{i_1\dots i_k}$ are given as the value $\Phi(e_{i_1}, \dots, e_{i_k})$ 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 Φ has vector values, it is given by finitely many components and the previous result can be used for each of them $\qquad \square$

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 4.1.24. 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*

$$(1) \qquad \frac{d}{dt}\langle u(t), v(t) \rangle = \langle u'(t), v(t) \rangle + \langle u(t), v'(t) \rangle$$

$$(2) \qquad \frac{d}{dt}(u(t) \times v(t)) = u'(t) \times v(t) + u(t) \times v'(t)$$

6.1.15. 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 $r(t)$ be a curve in the Euclidean space $\mathbb{R}^3$. For theoretical purposes, it is convenient to choose arc length s as a parameter. Since $dr^2 = dx^2 + dy^2 + dz^2 = ds^2$, 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.

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 unit vector $r'(s)$, the vector $r''(s)$, for which (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 $r'(s)$.

$$f''(x) > 0 \quad \text{pro} \quad x \to 0^+$$

implies that f is convex on the interval $[0, \pi/4)$, concave on $(\pi/4, \pi/2]$, convex on $[\pi/2, 3\pi/4)$ and concave on $(3\pi/4, \pi]$. The convexity and concavity of f on other subintervals is given by its periodicity and a simple observation: if a function is even and convex on an interval (a, b), where $0 \leq a < b$, then it's also convex on $(-b, -a)$.

All that's left is computing the derivative (to estimate the speed of the growrth of the function) at the point of inflection, yielding $f'(\pi/2) = 1$. Based on all previous results, it's now easy to plot the graph of function f. $\square$

6.A.36. Determine the course of the function

$$\frac{x}{\ln(x)},$$

and plot its graph.

Solution. i) First we'll determine the domain of the function: $\mathbb{R}^+ \setminus \{1\}$.

ii) We'll find the intervals of monotonicity of the function: first we'll find zero points of the derivative:

$$f'(x) = \frac{\ln(x) - 1}{\ln^2(x)} = 0$$

The root of this equation is e. Next we can see that $f'(x)$ is negative on both intervals $(0, 1)$ and $(1, e)$, hence $f(x)$ is decreasing on both intervals $(0, 1)$ and $(1, e)$. Additionally, $f'(x)$ is positive on the interval (e, ∞), thus $f(x)$ is increasing here. That means the function f has the only extreme at point e, being the minimum. (we can also decide this using the sign of the second derivative of the function f at point e, because $f^{(2)}(e) > 0$).

iii) We'll find the points of inflection:

$$f^{(2)}(x) = \frac{\ln(x) - 2}{x \ln^3(x)} = 0$$

The root of this equation is e^2, so it must be a point of inflection (it cannot be an extreme with regard to the ptrevious point).

iv) The asymptotes. The line $x = 1$ is an asymptote of the function. Next, let's look for asymptotes with a finite slope k:

$$k = lim_{x \to \infty} \frac{\frac{x}{\ln(x)}}{x} = \lim_{x \to \infty} \frac{1}{\ln(x)} = 0.$$

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.
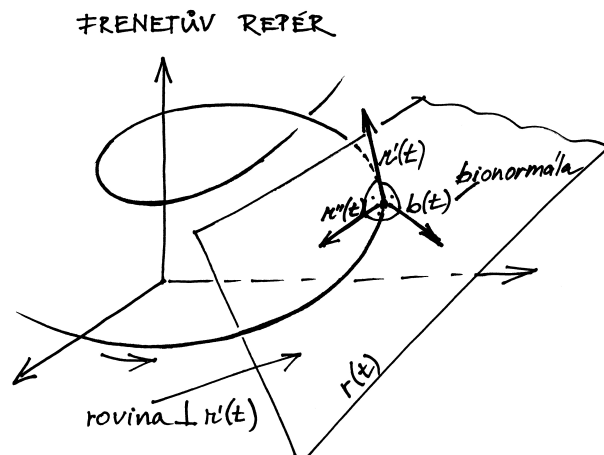
If the second derivative is nonzero, the normed vector

$$n(s) = \frac{1}{\|r''(s)\|} r''(s)$$

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

$$0 = \frac{d}{ds} \langle b(s), r'(s) \rangle = \langle b'(s), r'(s) \rangle + \langle b(s), r''(s) \rangle$$
$$= \langle b'(s, r'(s) \rangle + \kappa(s) \langle b(s), n(s) \rangle = \langle b'(s), r'(s) \rangle,$$

which shows that the derivative of the binormal is orthogonal to $r'(s)$. $b'(s)$ 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.

We have not yet computed the rate of change of the principal normal, which can be written as $n(s) = b(s) \times r'(s)$:

$$n'(s) = b'(s) \times r'(s) + \kappa(s) b(s) \times n(s)$$
$$= -\tau(s) n(s) \times r'(s) + \kappa(s) (-r'(s))$$
$$= \tau(s) b(s) - \kappa(s) r'(s).$$

Successively, for all points with nonzero second derivative of the curve $r(s)$ parametrized by the arc length, there is derived the important basis $(r'(s), n(s), b(s))$, called *the Frenet frame* in the classical literature. At the same time, this

If the asymptote exists, its slope must be 0. Let's continue the computation

$$\lim_{\times \to \infty} \frac{x}{\ln(x)} - 0 \cdot x = \lim_{x \to \infty} \ln(x) = \infty,$$

and because the limit isn't finite, an asymptote with a finite slope doesn't exist.

The course of the function:
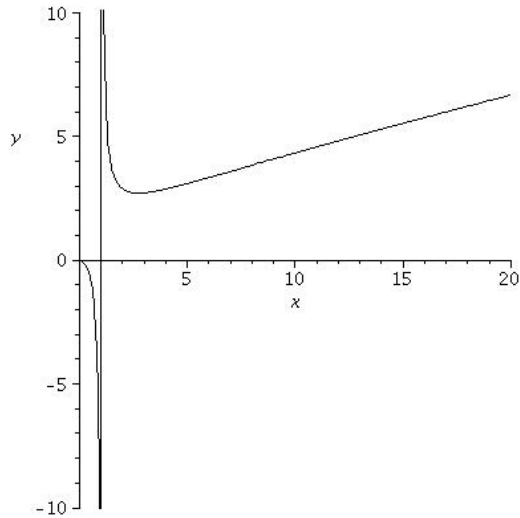
Now move from determining the course of functions onto other subjects connected to derivatives of functions. First we'll demonstrate the concept of curvature and the osculating circle on an ellipse

6.A.37. Determine the curvature of the ellipse $x^2 + 2y^2 = 2$ at its vertices (4.C.9). Also determine the equations of the circles of osculation at these vertices.

Solution. Because the ellipse is already in the basic form at the given coordinates (there are no mixed or linear terms), the given basis is already a polar basis. Its axes are the coordinate axes x and y, its vertices are the points $[\sqrt{2}, 0]$, $[-\sqrt{2}, 0]$, $[0, 1]$ and $[0, -1]$. Let's first compute the curvature at vertice $[0, 1]$. If we consider the coordinate y as a function of the coordinate x (determined uniquely in a neighbourhood of $[0, 1]$), then differentiating the equation of the ellipse with respect to the variable x yields $2x + 4yy' = 0$, hence $y' = -\frac{x}{2y}$ (y' denotes the derivative of function $y(x)$ with respect to the variable x; in fact it's nothing else than expressing the derivative of a function given implicitly, see ??). Differentiating this

basis is used in order to express the derivatives of its components in the form of the *Frenet–Serret formulas*

$$\frac{dr'}{ds}(s) = \kappa(s)n(s), \quad \frac{dn}{ds}(s) = \tau(s)b(s) - \kappa(s)r'(s)$$

$$\frac{db}{ds}(s) = -\tau(s)n(s).$$

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 a space parametrized by the length of their arc can be mapped to each other by an Euclidean transformation if and only if their curvature functions and torsion functions coincide except for a constant shift of the parameter. Moreover, for every choice of smooth functions κ a τ 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 κ 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 $x(t)$, $y(t) = f(x(t))$):

$$dt = (1 + (f_x)^2)^{1/2}dx, \quad dx = (1 + (f_x)^2)^{-1/2}dt$$

(here we write $f_x = \frac{df}{dx}$) we obtain the following equality for the unit tangent vector of the graph of a curve

$$r'(s) = (x'(s), y'(s)) = \left((1 + (f_x)^2)^{-1/2}, f_x(1 + (f_x)^2)^{-1/2}\right).$$

A messy, but very similar computation for the second derivative and its length leads to

$$\kappa^2 = \|r''\|^2 = \left(\frac{d^2 f}{dx^2}\right)^2 (1 + (f_x)^2)^{-3}$$

as expected. If we write $r = (x, y)$, $y' = f_x x'$, $x' = (1 + f_x^2)^{-1/2}$, then

$$x'' = -\frac{1}{2}(x')^3 2f_x f_{xx} x' = -(x')^4 f_x f_{xx}$$
$$y'' = f_{xx}(x')^2 + f_x x'' = f_{xx}(x')^2 - f_{xx} f_x^2 (x')^4.$$

Hence

$$(x'')^2 + (y'')^2 = f_{xx}^2 (x')^8 \left(f_x + (1 + f_x^2)^2 + f_x^4\right.$$
$$\left. - 2f_x^2(1 + f_x^2)\right)$$
$$= f_{xx}^2 (1 + f_x^2)^{-4}(f_x^2 + 1)$$
$$= f_{xx}^2 (1 + f_x^2)^{-3}.$$

6.1.16. The numerical derivatives. In the begining 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.

389

equation with respect to x than yields $y'' = -\frac{1}{2}(\frac{1}{y} - \frac{xy'}{y^2})$. At point $[1, 0]$, we obtain $y' = 0$ and $y'' = -\frac{1}{2}$ (we'd receive the same results if we explicitly expressed $y = \frac{1}{2}\sqrt{2 - x^2}$ from the equation of the ellipse and performed differentiation; the computation would be only a little more complicated, as the reader can surely verify). According to 6.1.12, the radius of the osculation circle will be

$$\frac{(1 + (y')^2)^{\frac{3}{2}}}{(y'')^2} = -2,$$

or 2, respectively, and the sign tells us the circle will be "below" the graph of the function. The ideas in 6.1.12 and 6.1.15 imply that its center will be in the direction opposite to the normal line of this curve, i.e. on the y axis (the function y as a function of variable x has a derivative at point $[0, 1]$, thus the tangent line to its graph at this point will be parallel to the x axis, and because the normal is perpendicular to the tangent, it must be the y axis at this point). The radius is 2, so the center will be at point $[0, 1 - 2] = [0, -1]$. In total, the equation of the osculation circle of the ellipse $x^2 + 2y^2 = 2$ at point $[0, 1]$ will be $x^2 + (y + 1)^2 = 4$. Analogously, we can determine the equation of the osculation circle at point $[0, -1]$: $x^2 + (y - 1)^2 = 4$. The curvatures of the ellipse (as a curve) at these points then equal $\frac{1}{2}$ (the absolute value of the curvature of the graph of the function).

For determining the osculation circle at 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 (in a neighbourhood of point $[\sqrt{2}, 0]$, the variable y as a function of x isn't determined uniquely, so we cannot use the previous procedure - technically it would end up by diving by zero). Sequentially, we obtain: $2xx' + 4y = 0$, thus $x' = -2\frac{y}{x}$, and $x'' = -2(\frac{1}{x} - \frac{yx'}{x^2})$. 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 6.1.12. 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 $\sqrt{2}$.

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 $x_0 = a, x_1, x_2, \ldots, x_n = b$, are given while $x_i - x_{i-1} = h$ for some constant $h > 0$ and all indices $i = 1, \ldots, n$. Write the Taylor expansion of function f in the form

$$f(x_i \pm h) = f_i \pm hf'(x_i) + \frac{h^2}{2}f''(x_i) \pm \frac{h^3}{3!}f^{(3)}(x_i) + \ldots$$

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 approximated 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*.

ASYMPTOTIC ESTIMATES

Definition. The expression $G(h)$ is asymptotically equal to $F(h)$ for $h \to 0$. Write $G(h) = O(F(h))$, if the finite limit

$$\lim_{h \to 0} \frac{G(h)}{F(h)} = a \in \mathbb{R}$$

exists.

Similarly, compare the expressions for $h \to \infty$ and use the same notation.

Denote the values of the derivatives of $f(x)$ at the points x_i as $f_i^{(j)}$. Write the Taylor expansion as:

$$f_{i\pm1} = f_i \pm f_i'h + \frac{f_i''}{2}h^2 \pm \frac{f_i'''}{6}h^3 + \ldots$$

Considering combinations of the two expansions and f_i itself, we can express the derivative f_i' as follows

$$\frac{f_{i+1} - f_{i-1}}{2h} = f_i' + \frac{h^2}{3!}f_i^{(3)} + \ldots$$
$$\frac{f_{i+1} - f_i}{h} = f_i' + \frac{h}{2!}f_i'' + \ldots$$
$$\frac{f_i - f_{i-1}}{h} = f_i' - \frac{h}{2!}f_i'' + \ldots$$

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 $f_i' = \frac{f_i - f_{i-1}}{h}$.

If we use the Taylor expansions with remainder of the appropriate order, we obtain an expression of the error of the

□
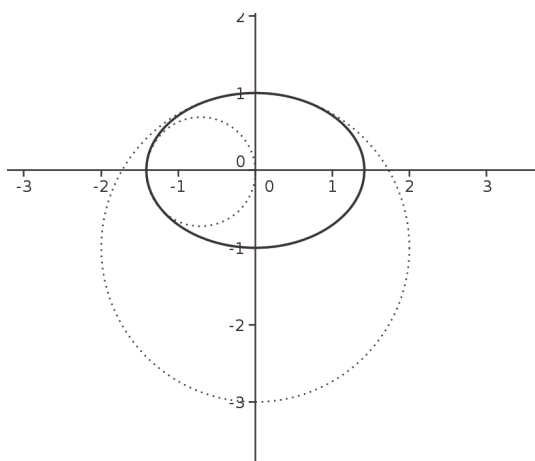
6.A.38. Remark. The vertices of an ellipse (more generally the vertices of a closed smooth curve in plane) can be defined as the points at which the function of curvature has an extreme. The ellipse having four vertices isn't a coincidence. The so called "Four vertices theorem" states that a closed curve 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 $\sqrt{2}$.

B. Integration

We start with an example testing the understanding the concept of Riemannian integration.

6.B.1. Let $y = |x|$ on the interval $I = [-1, 1]$ and let
$$\Xi_n = \left(-1, -\frac{n-1}{n}, \dots, -\frac{1}{n}, 0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\right)$$
be a partition of the interval I for arbitrary $n \in \mathbb{N}$. Determine $S_{\Xi_n, \sup}$ and $S_{\Xi_n, \inf}$ (the upper and lower Riemann sum corresponding to the given partition).

Based on this result decide if the function $y = |x|$ on $[-1, 1]$ is integrable (in Riemann sense). ○

And now some easy examples that everyone should handle.

6.B.2. Using integration "by heart", express

(a) $\int e^{-x} \, dx$, $x \in \mathbb{R}$;

(b) $\int \frac{1}{\sqrt{4-x^2}} \, dx$, $x \in (-2, 2)$;

(c) $\int \frac{1}{x^2+3} \, dx$, $x \in \mathbb{R}$;

(d) $\int \frac{3x^2+1}{x^3+x+2} \, dx$, $x \neq -1$.

Solution. We can easily obtain

approximation by the central difference in the form
$$\frac{1}{3!}h^2\left(f^{(3)}(x_i + \xi h) - f^{(3)}(x_i - \eta h)\right).$$

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. Thus, under the assumption of bounded derivatives of third or second order, the asymptotic estimates are computed:.

Theorem. *The asymptotic estimate of the error of the central difference is $O(h^2)$. The errors of the backward and forward differences are $O(h)$.*

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.

2. Integration

6.2.1. Indefinite integral. Now, we reverse the procedure of differention. 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. 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,
$$F(x) - G(x) = F(a) - G(a) + 0 \cdot (x - a).$$

(a) $\int e^{-x}\,dx = -\int -e^{-x}\,dx = -e^{-x} + C$;

(b) $\int \frac{1}{\sqrt{4-x^2}}\,dx = \int \frac{\frac{1}{2}}{\sqrt{1-\left(\frac{x}{2}\right)^2}}\,dx = \arcsin \frac{x}{2} + C$;

(c) $\int \frac{1}{x^2+3}\,dx = \frac{1}{3}\int \frac{1}{\frac{x^2}{3}+1}\,dx = \frac{1}{\sqrt{3}}\int \frac{\frac{1}{\sqrt{3}}}{1+\left(\frac{x}{\sqrt{3}}\right)^2}\,dx = \frac{1}{\sqrt{3}}\operatorname{arctg}\frac{x}{\sqrt{3}} + C$;

(d) $\int \frac{3x^2+3}{x^3+3x+2}\,dx = \ln\left|x^3+3x+2\right| + C$,

where we used the formula $\int \frac{f'(x)}{f(x)}\,dx = \ln|f(x)| + C$.

□

6.B.3. Compute the indefinite integral

$\int \left(7^x + 4\,e^{\frac{2x}{3}} - \frac{1}{2^x} + 9\sin 5x + 2\cos\frac{x}{2} - \frac{3}{\cos^2 x} + \frac{1}{3-x}\right)dx$

for $x \neq 3$, $x \neq \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$.

Solution. Only by combining the earlier derived formulas, we obtain

$\int \left(7^x + 4\,e^{\frac{2x}{3}} - \frac{1}{2^x} + 9\sin 5x + 2\cos\frac{x}{2} - \frac{3}{\cos^2 x} + \frac{1}{3-x}\right)dx$
$\frac{7^x}{\ln 7} + 6\,e^{\frac{2x}{3}} + \frac{1}{2^x \ln 2} - \frac{9}{5}\cos 5x + 4\sin\frac{x}{2} - 3\operatorname{tg} x - \ln|3-x| + C.$

□

For expressing the following integrals, we'll use the method of integration by parts (see 6.2.3).

6.B.4. Compute $\int x\cos x\;dx$, $x \in \mathbb{R}$ and $\int \ln x\,dx$, $x > 0$;

Solution.

$\int \ln x\,dx = \begin{vmatrix} u = \ln x & u' = \frac{1}{x} \\ v' = 1 & v = x \end{vmatrix}$
$= x\ln x - \int 1\,dx = x\ln x - x + C.$

$\int x\cos x\;dx = \begin{vmatrix} u = x & u' = 1 \\ v' = \cos x & v = \sin x \end{vmatrix}$
$= x\sin x - \int \sin x\,dx = x\sin x + \cos x + C.$

□

6.B.5. Using integration by parts, compute

(a) $\int \left(x^2 + 1\right)e^{-x}\,dx$, $x \in \mathbb{R}$,

(b) $\int (2x - 1)\ln x\;dx$, $x > 0$,

(c) $\int \operatorname{arctg} x\;dx$, $x \in \mathbb{R}$,

(d) $\int e^x \sin(\beta x)\;dx$, $x, \beta \in \mathbb{R}$,

Thus the difference of the values of the functions F and G is constant on the interval $[a, b]$.

□

The previous lemma supports another notation for the indefinite integral:

$$F(x) = \int f(x)\,dx + C$$

with an unknown constant C.

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 a 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

$$a = x_0 < x_1 < \cdots < x_n = b.$$

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:

$$\sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i) \simeq \sum_{i=0}^{n-1} \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i}(x_{i+1} - x_i)$$
$$= \sum_{i=0}^{n-1} (F(x_{i+1}) - F(x_i)) = F(b) - F(a).$$

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*.[3]

_____

[3]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.

Solution. First emphasise that by integration by parts, we can compute every integral in the form of

$$\int P(x)\, a^{bx}\, dx, \quad \int P(x) \sin(bx)\, dx, \quad \int P(x) \cos(bx)\, dx,$$

$$\int P(x) \log_a^n x\, dx, \qquad \int x^b \log_a^n (kx)\, dx,$$

$$\int P(x) \arcsin(bx)\, dx, \qquad \int P(x) \arccos(bx)\, dx,$$

$$\int P(x) \operatorname{arctg}(bx)\, dx, \qquad \int P(x) \operatorname{arccotg}(bx)\, dx,$$

$$\int a^{bx} \sin(cx)\, dx, \qquad \int a^{bx} \cos(cx)\, dx,$$

where P is an arbitrary polynomial and

$$a \in (0,1) \cup (1, +\infty), \quad b, c \in \mathbb{R} \setminus \{0\}, \quad n \in \mathbb{N}, \quad k > 0.$$

Thus we know that

(a)

$$\int \left(x^2 + 1\right) e^{-x}\, dx =$$
$$\left|\begin{array}{c|c} F(x) = x^2 + 1 & F'(x) = 2x \\ G'(x) = e^{-x} & G(x) = -e^{-x} \end{array}\right| =$$
$$- \left(x^2 + 1\right) e^{-x} + \int 2x\, e^{-x}\, dx =$$
$$\left|\begin{array}{c|c} F(x) = 2x & F'(x) = 2 \\ G'(x) = e^{-x} & G(x) = -e^{-x} \end{array}\right| =$$
$$- \left(x^2 + 1\right) e^{-x} - 2x\, e^{-x} + \int 2\, e^{-x}\, dx =$$
$$- \left(x^2 + 1\right) e^{-x} - 2x\, e^{-x} - 2\, e^{-x} + C =$$
$$-e^{-x} \left(x^2 + 2x + 3\right) + C;$$

(b)

$$\int (2x - 1) \ln x\, dx =$$
$$\left|\begin{array}{c|c} F(x) = \ln x & F'(x) = 1/x \\ G'(x) = 2x - 1 & G(x) = x^2 - x \end{array}\right| =$$
$$\left(x^2 - x\right) \ln x - \int \frac{x^2 - x}{x}\, dx =$$
$$\left(x^2 - x\right) \ln x + \int 1 - x\, dx = \left(x^2 - x\right) \ln x + x - \frac{x^2}{2} + C;$$

(c)

$$\int \operatorname{arctg} x\, dx = \left|\begin{array}{c|c} F(x) = \operatorname{arctg} x & F'(x) = \frac{1}{1+x^2} \\ G'(x) = 1 & G(x) = x \end{array}\right| =$$
$$x \operatorname{arctg} x - \int \frac{x}{1+x^2}\, dx = x \operatorname{arctg} x - \frac{1}{2} \int \frac{2x}{1+x^2}\, dx =$$
$$x \operatorname{arctg} x - \frac{1}{2} \ln \left(1 + x^2\right) + C;$$

(d)

$$\int e^x \sin(\beta x)\, dx =$$
$$\left|\begin{array}{c|c} F(x) = e^x & F'(x) = e^x \\ G'(x) = \sin(\beta x) & G(x) = -\frac{1}{\beta} \cos \beta x \end{array}\right| =$$
$$-\frac{1}{\beta} e^x \cos(\beta x) + \frac{1}{\beta} \int e^x \cos(\beta x)\, dx =$$
$$\left|\begin{array}{c|c} F(x) = e^x & F'(x) = e^x \\ G'(x) = \cos(\beta x) & G(x) = \frac{1}{\beta} \sin(\beta x) \end{array}\right| =$$
$$-\frac{1}{\beta} e^x \cos(\beta x) + \frac{1}{\beta^2} e^x \sin()\beta x) - \frac{1}{\beta^2} \int e^x \sin(\beta x)\, dx,$$

which implies

$$\int e^x \sin x\, dx = \frac{1}{1+\beta^2}\, e^x \left(\sin(\beta x) - \beta \cos(\beta x)\right) + C.$$

□

$$F'(x) = f(x)$$
$$f(x_0) \approx \frac{F(x_2) - F(x_1)}{x_2 - x_1}$$
$$f(b)$$
$$f(a)$$
$$a \quad x_1 \quad x_2 \quad x_3 \quad b$$
plocha $\simeq F(x_2) - F(x_1)$
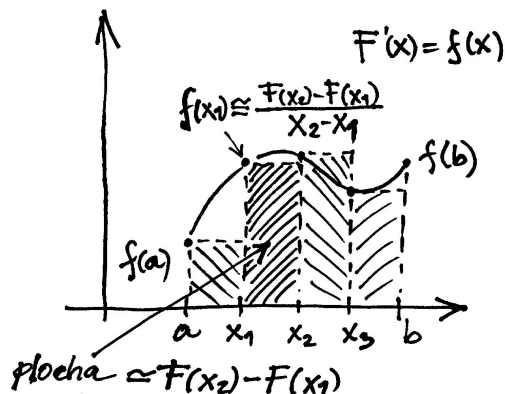
NEWTON INTEGRAL

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.

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 sequel.

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:

For expressing the following integrals, it's convenient to use the substitution method (see 6.2.5).

6.B.6. Using a suitable substitution, determine

(a) $\int \sqrt{2x-5}\,dx$, $x > \frac{5}{2}$;

(b) $\int \frac{(7+\ln x)^7}{x}\,dx$, $x > 0$;

(c) $\int \frac{\cos x}{(1+\sin x)^2}\,dx$, $x \neq \frac{(3+4k)\pi}{2}$, $k \in \mathbb{Z}$;

(d) $\int \frac{\cos x}{\sqrt{1+\sin^2 x}}\,dx$, $x \in \mathbb{R}$.

Solution. We have

(a)

$$\int \sqrt{2x-5}\,dx = \left| \begin{array}{l} t = 2x-5 \\ dt = 2\,dx \end{array} \right| = \frac{1}{2}\int \sqrt{t}\,dt$$
$$= \frac{1}{3} t^{\frac{3}{2}} + C = \frac{1}{3}\sqrt{(2x-5)^3} + C;$$

(b)

$$\int \frac{(7+\ln x)^7}{x}\,dx = \left| \begin{array}{l} t = 7+\ln x \\ dt = \frac{1}{x}\,dx \end{array} \right| = \int t^7\,dt = \frac{t^8}{8} + C$$
$$= \frac{(7+\ln x)^8}{8} + C;$$

(c)

$$\int \frac{\cos x}{(1+\sin x)^2}\,dx = \left| \begin{array}{l} t = 1+\sin x \\ dt = \cos x\,dx \end{array} \right| = \int \frac{dt}{t^2}$$
$$= -\frac{1}{t} + C = -\frac{1}{1+\sin x} + C;$$

(d)

$$\int \frac{\cos x}{\sqrt{1+\sin^2 x}}\,dx = \left| \begin{array}{l} t = \sin x \\ dt = \cos x\,dx \end{array} \right| = \int \frac{1}{\sqrt{1+t^2}}\,dt$$
$$= \left| \begin{array}{l} u = t + \sqrt{1+t^2} > 0 \\ du = \left(1 + \frac{t}{\sqrt{1+t^2}}\right)dt \\ \frac{du}{t+\sqrt{1+t^2}} = \frac{1}{\sqrt{1+t^2}}\,dt \end{array} \right| = \int \frac{1}{u}\,du = \ln u + C =$$
$$\ln\left(t + \sqrt{1+t^2}\right) + C$$
$$= \ln\left(\sin x + \sqrt{1+\sin^2 x}\right) + C. \qquad \square$$

6.B.7. Determine the integrals

a) $\int \frac{dx}{\sin^2(x) - \cos^2(x)}$,

b) $\int x^2 \sqrt{2x+1}\,dx$.

INTEGRATION TABLE

For arbitrary nonzero $a, b \in \mathbb{R}$ and $n \in \mathbb{Z}$, $n \neq -1$:

$$\int a\,dx = ax + C$$

$$\int ax^n\,dx = \frac{a}{n+1}x^{n+1} + C$$

$$\int e^{ax}\,dx = \frac{1}{a}e^{ax} + C$$

$$\int \frac{a}{x}\,dx = a\ln x + C$$

$$\int a\cos(bx)\,dx = \frac{a}{b}\sin(bx) + C$$

$$\int a\sin(bx)\,dx = -\frac{a}{b}\cos(bx) + C$$

$$\int a\cos(bx)\sin^n(bx)\,dx = \frac{a}{b(n+1)}\sin^{n+1}(bx) + C$$

$$\int a\sin(bx)\cos^n(bx)\,dx = -\frac{a}{b(n+1)}\cos^{n+1}(bx) + C$$

$$\int a\,\mathrm{tg}(bx)\,dx = -\frac{a}{b}\ln(\cos(bx)) + C$$

$$\int \frac{a}{a^2+x^2}\,dx = \mathrm{arctg}\left(\frac{x}{a}\right) + C$$

$$\int \frac{-1}{\sqrt{a^2-x^2}}\,dx = \arccos\left(\frac{x}{a}\right) + C$$

$$\int \frac{1}{\sqrt{a^2-x^2}}\,dx = \arcsin\left(\frac{x}{a}\right) + C.$$

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.

Further rules can be added by observations of suitable 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 integral 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.

Solution. For computing the first integral, we'll choose the substitution $t = \operatorname{tg} x$, which can be often used with an advantage.

$$\int \frac{\mathrm{d}x}{\sin^2(x) - \cos^2(x)}$$

$$= \left| \begin{array}{l} \text{substitution } t = \operatorname{tg} x \\ dt = \frac{1}{\cos^2 x}\,\mathrm{d}x = (1 + \operatorname{tg}^2(x))\,\mathrm{d}x = (1 + t^2)\,\mathrm{d}x \\ \sin^2(x) = \frac{\operatorname{tg}^2(x)}{1 + \operatorname{tg}^2(x)} = \frac{t^2}{1 + t^2} \\ \cos^2(x) = \frac{1}{1 + \operatorname{tg}^2(x)} = \frac{1}{1 + t^2} \end{array} \right|$$

$$= \int \frac{1}{t^2 - 1}\,\mathrm{d}t = \frac{1}{2} \int \frac{1}{t - 1} - \frac{1}{2} \int \frac{1}{t + 1}$$

$$= \frac{1}{2} \ln\left(\frac{\operatorname{tg}(x) - 1}{\operatorname{tg} + 1} \right) + C$$

Now we'll compute the second integral:

$$\int x^2 \sqrt{2x + 1}\,\mathrm{d}x$$

$$= \left| \begin{array}{ll} u = x^2 & u = 2x \\ v' = \sqrt{2x + 1} & v = \frac{1}{3}(2x + 1) \end{array} \right|$$

$$= \frac{1}{3} x^2 (2x + 1)^{\frac{3}{2}} - \frac{4}{3} \int x^2 \sqrt{2x + 1}\,\mathrm{d}x - \frac{2}{9}(2x + 1)^{\frac{3}{2}} + C,$$

which can be thought of as an equation, when the variable is the integral. By putting it on one side,

$$\int x^2 \sqrt{2x + 1}\,\mathrm{d}x$$

$$= \frac{1}{7} x^2 (2x + 1)^{\frac{3}{2}} - \frac{2}{7} \int x \sqrt{2x + 1}$$

$$\left| \begin{array}{ll} u = x & u' = 1 \\ v' = \sqrt{2x + 1} & v = \frac{1}{3}\sqrt{2x + 1} \end{array} \right|$$

$$= \frac{1}{7} x^2 (2x + 1)^{\frac{3}{2}} - \frac{2}{7} \left(\frac{1}{3} x \sqrt{2x + 1} - \frac{1}{3} \int (2x + 1)^{\frac{3}{2}}\,\mathrm{d}x \right)$$

$$= \frac{1}{7} x^2 (2x + 1)^{\frac{3}{2}} - \frac{2}{21} x \sqrt{2x + 1} + \frac{2}{105}(2x + 1)^{\frac{5}{2}}$$

$$= \frac{1}{7} x^2 (2x + 1)^{\frac{3}{2}} - \frac{2}{35} x(2x + 1)^{\frac{3}{2}} + \frac{2}{105}(2x + 1)^{\frac{3}{2}} + C$$

$\square$

6.B.8. Using the basic formulas, compute

(a) $\int \frac{1}{\sqrt[3]{x}}\,dx$, $x \neq 0$;

(b) $\int \operatorname{tg}^2 x\,dx$, $x \neq \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$;

(c) $\int \frac{\cos x}{1 + \sin x}\,dx$, $x \neq -\frac{\pi}{2} + 2k\pi$, $k \in \mathbb{Z}$;

(d) $\int 6 \sin 5x + \cos \frac{x}{2} + 2\,e^{\frac{2x}{3}}\,dx$, $x \in \mathbb{R}$.

Solution. Case (a). We can immediately determine

$$\int \frac{1}{\sqrt[3]{x}}\,dx = \int x^{-1/3}\,dx = \frac{x^{\frac{2}{3}}}{\frac{2}{3}} + C = \frac{3}{2}\sqrt[3]{x^2} + C,$$

where the notation in which we add $C \in \mathbb{R}$ has to be understood in a way that we can get all primitive functions exactly by a constant translation of an arbitrary primitive function.

The formula for computing the integral on the left hand side

$$\int F(x)G'(x)\,dx = F(x) \cdot G(x) - \int F'(x)G(x)\,dx + C$$

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 $G'(x) = 1$:

$$\int \ln x\,dx = \int 1 \cdot \ln x\,dx$$

$$= x \ln x - \int \frac{1}{x} x\,dx = x \ln x - x + C.$$

6.2.5. Integration by substitution. Another useful procedure is derived from the chain rule for differentiating composite functions. If

$$F'(y) = f(y), \qquad y = \varphi(x),$$

where φ 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

$$F(\varphi(x)) + C = \int f(\varphi(x))\varphi'(x)\,dx.$$

By substituting $x = \varphi^{-1}(y)$, we obtain the originally desired primitive function. This is often written as follows:

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

$$f(y)\,dy = f(\varphi(x))\varphi'(x)\,dx.$$

But that's only true on an interval. In other words, the value C is generally distinct for $x < 0$ and for $x > 0$. Thus we should consider the values C_1 and C_2. For the sake of simplicity though, we'll use the notation without indices and stating the corresponding intervals. Furthermore, we'll help ourselves by letting $aC = C$ for $a \in \mathbb{R} \smallsetminus \{0\}$ and $C + b = C$ for $b \in \mathbb{R}$, based on the fact that

$$\{C; \, C \in \mathbb{R}\} = \{aC; \, C \in \mathbb{R}\} = \{C + b; \, C \in \mathbb{R}\} = \mathbb{R}.$$

We could then obtain an entirely correct expression for example by substitutions $\hat{C} = aC$, $\tilde{C} = C + b$. These simplifications will prove their usefulness when computing more complicated problems, because they make the procedures and the simplifications more lucid.

Case (b). Sequential simplifications of the integrated function lead to

$$\int \operatorname{tg}^2 x \, dx = \int \frac{\sin^2 x}{\cos^2 x} \, dx = \int \frac{1 - \cos^2 x}{\cos^2 x} \, dx = \int \frac{1}{\cos^2 x} \, dx - \int 1 \, dx = \operatorname{tg} x - x + C,$$

where we helped ourselves by the knowledge of the derivative

$$(\operatorname{tg} x)' = \frac{1}{\cos^2 x}, \qquad x \neq \frac{\pi}{2} + k\pi, \, k \in \mathbb{Z}.$$

Case (c). It suffices to realize that this is a special case of the formula

$$\int \frac{f'(x)}{f(x)} \, dx = \ln |f(x)| + C,$$

which can be verified directly by differentiation

$$(\ln|f(x)| + C)' = (\ln [\pm f(x)])' + (C)' = \frac{[\pm f(x)]'}{\pm f(x)} = \frac{\pm f'(x)}{\pm f(x)} = \frac{f'(x)}{f(x)}.$$

Hence

$$\int \frac{\cos x}{1 + \sin x} \, dx = \ln (1 + \sin x) + C.$$

Case (d). Because the integral of a sum is the sum of integrals (if the seperate integrals are sensible) and a nonzero constant can be factored out of the integral at any time, we have

$$\int 6 \sin 5x + \cos \tfrac{x}{2} + 2 \, e^{\frac{2x}{3}} \, dx = -\tfrac{6}{5} \cos 5x + 2 \sin \tfrac{x}{2} + 3 \, e^{\frac{2x}{3}} + C.$$

$\square$

6.B.9. Determine

(a) $\int \frac{x}{\cos^2 x} \, dx$, $x \neq \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$;

(b) $\int x^2 \, e^{-3x} \, dx$, $x \in \mathbb{R}$;

(c) $\int \cos^2 x \, dx$, $x \in \mathbb{R}$.

Solution. Case (a). Using integration by parts, we obtain

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 $x = \sin t$. Then $dx = \cos t \, dt$. So

$$I = \int \frac{1}{\sqrt{1 - \sin^2 t}} \cos t \, dt = \int \frac{1}{\sqrt{\cos^2 t}} \cos t \, dt$$

$$= \int dt = t + C.$$

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 recurences. Often the use of substitutions and integrating by parts leads to recurent relations, from which desired integrals can be evaluated. We illustrate by an example. Integrating by parts, to evaluate

$$I_m = \int \cos^m x \, dx = \int \cos^{m-1} x \cos x \, dx$$

$$= \cos^{m-1} x \sin x - (m - 1) \int \cos^{m-2} x (-\sin x) \sin x \, dx$$

$$= \cos^{m-1} x \sin x + (m - 1) \int \cos^{m-2} x \sin^2 x \, dx.$$

Using the formula $\sin^2 x = 1 - \cos^2 x$,

$$m I_m = \cos^{m-1} x \sin x + (m - 1) I_{m-2}$$

The initial values are

$$I_0 = x, \quad I_1 = \sin x.$$

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

$\int \frac{x}{\cos^2 x}\, dx = \begin{vmatrix} F(x) = x & F'(x) = 1 \\ G'(x) = \frac{1}{\cos^2 x} & G(x) = \operatorname{tg} x \end{vmatrix} = x \operatorname{tg} x -$
$\int \operatorname{tg} x\, dx = x \operatorname{tg} x + \int \frac{-\sin x}{\cos x}\, dx = x \operatorname{tg} x + \ln|\cos x| + C.$

Case (b). This time we are clearly integrating a product of two functions. By applying the method of integration by parts, we reduce the integral to another integral in a way that we differentiate one function and integrate the second. We can integrate both of them (we can differentiate all elementary functions). Thus we must decide which of the two variants of the method we'll use (whether we'll integrate the function $y = x^2$, or $y = \mathrm{e}^{-3x}$). Notice that we can use integration bz parts repeatedlz and that the n-th derivative of a polynomial of degree $n \in \mathbb{N}$ is a constant polynomial. That gives us a way to compute

$\int x^2\, \mathrm{e}^{-3x}\, dx = \begin{vmatrix} F(x) = x^2 & F'(x) = 2x \\ G'(x) = \mathrm{e}^{-3x} & G(x) = -\frac{1}{3}\,\mathrm{e}^{-3x} \end{vmatrix} =$
$-\frac{1}{3}\, x^2\, \mathrm{e}^{-3x} + \frac{2}{3} \int x\, \mathrm{e}^{-3x}\, dx$

and furthermore

$\int x\, \mathrm{e}^{-3x}\, dx = \begin{vmatrix} F(x) = x & F'(x) = 1 \\ G'(x) = \mathrm{e}^{-3x} & G(x) = -\frac{1}{3}\,\mathrm{e}^{-3x} \end{vmatrix} =$
$-\frac{1}{3}\, x\, \mathrm{e}^{-3x} + \frac{1}{3} \int \mathrm{e}^{-3x}\, dx = -\frac{1}{3}\, x\, \mathrm{e}^{-3x} - \frac{1}{9}\, \mathrm{e}^{-3x} + C.$

In total, we have

$\int x^2\, \mathrm{e}^{-3x}\, dx = -\frac{1}{3}\, x^2\, \mathrm{e}^{-3x} - \frac{2}{9}\, x\, \mathrm{e}^{-3x} - \frac{2}{27}\, \mathrm{e}^{-3x} + C =$
$-\frac{1}{3} \mathrm{e}^{-3x} \left(x^2 + \frac{2}{3} x + \frac{2}{9}\right) + C.$

Note that a repeated use of integration by parts within the scope of computing one integral is common (just like when computeing limits by the l'Hospital rule).

Case (c). Again we apply integration by parts using

$\int \cos^2 x\, dx = \int \cos x \cdot \cos x\, dx =$
$\begin{vmatrix} F(x) = \cos x & F'(x) = -\sin x \\ G'(x) = \cos x & G(x) = \sin x \end{vmatrix} =$
$\cos x \cdot \sin x + \int \sin^2 x\, dx = \cos x \cdot \sin x + \int 1 - \cos^2 x\, dx =$
$\cos x \cdot \sin x + \int 1\, dx - \int \cos^2 x\, dx =$
$\cos x \cdot \sin x + x - \int \cos^2 x\, dx.$

Although the return to the given integral might make the reader cast some doubts on it, the equality

$\int \cos^2 x\, dx = \cos x \cdot \sin x + x - \int \cos^2 x\, dx$

implies

$2 \int \cos^2 x\, dx = \cos x \cdot \sin x + x + C,$

i.e.

(1) $\qquad \int \cos^2 x\, dx = \frac{1}{2}\, (x + \sin x \cdot \cos x) + C.$

It suffices to remember that we put $C/2 = C$ and that the indefinite integral (as an infinite set) can be represented by one specific function and its translations.

example while integrating over the interval $[0, 2\pi]$, the integrals have these values:

$I_0 = \int_0^{2\pi} dx = [x]_0^{2\pi} = 2\pi$

$I_1 = \int_0^{2\pi} \cos x\, dx = [\sin x]_0^{2\pi} = 0$

$I_m = \int_0^{2\pi} \cos^m x\, dx = \begin{cases} 0 & \text{for even } m \\ \frac{m-1}{m} I_{m-2} & \text{for odd } m \end{cases}.$

Thus for even $m = 2n$, the result is

$\int_0^{2\pi} \cos^{2n} x\, dx = \frac{(2n-1)(2n-3)\ldots 3 \cdot 1}{2n(2n-2)\ldots 2}\, 2\pi.$

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

$$f = q \cdot g + h, \qquad \frac{f}{g} = q + \frac{h}{g}.$$

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*).

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:

$$\frac{4x + 2}{x^2 + 3x + 2} = \frac{4x + 2}{(x + 1)(x + 2)} = \frac{A}{x + 1} + \frac{B}{x + 2}.$$

We emphasise that usually suitable simplifications or substitutions lead to the result faster than integration by parts. For example, by using the identity

$$\cos^2 x = \tfrac{1}{2}\left(1 + \cos 2x\right), \qquad x \in \mathbb{R}$$

we easily obtain

$\int \cos^2 x \, dx = \int \tfrac{1}{2}\, dx + \int \tfrac{1}{2}\cos 2x\, dx = \tfrac{x}{2} + \tfrac{\sin 2x}{4} + C = \tfrac{x}{2} + \tfrac{2\sin x\cos x}{4} + C = \tfrac{1}{2}\left(x + \sin x \cdot \cos x\right) + C.$

$\square$

6.B.10. Integrate

(a) $\int \cos^5 x \cdot \sin x \, dx,\ x \in \mathbb{R}$;

(b) $\int \cos^5 x \cdot \sin^2 x \, dx,\ x \in \mathbb{R}$;

(c) $\int \frac{\sin^4 x}{\cos^4 x} \, dx,\ x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$;

(d) $\frac{1 - \sqrt[3]{x} + \sqrt{x}}{\sqrt[6]{x^5} + x} \, dx,\ x > 0$.

Solution. Case (a). This is a simple problem for the so called first substitution method, whose essence is writing the integral in the form of

(1) $$\int f(\varphi(x))\, \varphi'(x) \, dx$$

for certaing functions f and φ. Using the substitution $y = \varphi(x)$, (we also substitute $dy = \varphi'(x)\, dx$, which we get by differentiating $y = \varphi(x)$) , such integral can be reduced to the integral $\int f(y)\, dy$. By substituing $y = \cos x$, where $dy = -\sin x \, dx$, we then obtain

$$\int \cos^5 x \cdot \sin x \, dx = -\int \cos^5 x \, (-\sin x)\ dx =$$
$$-\int y^5 \, dy =$$
$$-\frac{y^6}{6} + C = -\frac{\cos^6 x}{6} + C.$$

Case (b). Using the equality

$$\int \cos^5 x \cdot \sin^2 x \, dx = \int \left(\cos^2 x\right)^2 \sin^2 x \cdot \cos x \, dx =$$
$$\int \left(1 - \sin^2 x\right)^2 \sin^2 x \cdot \cos x \, dx$$

we're tempted to use the substitution $t = \sin x$, which yields

$$\int \cos^5 x \cdot \sin^2 x \, dx = \left| \begin{array}{c} t = \sin x \\ dt = \cos x \, dx \end{array} \right| =$$
$$\int \left(1 - t^2\right)^2 t^2 \, dt =$$
$$\int t^6 - 2t^4 + t^2 \, dt = \frac{t^7}{7} - 2\frac{t^5}{5} + \frac{t^3}{3} + C =$$
$$\frac{\sin^7 x}{7} - \frac{2\sin^5 x}{5} + \frac{\sin^3 x}{3} + C.$$

Case (c). Because both sine and cosine are contained in an even power, we cannot proceed as in the previous problem. Let's try to use the so called second substitution method, which means a reduction of $\int f(y)\, dy$ to the form (1) for $y = \varphi(x)$. A situation in which we replace a simple expression by a more complicated one might seem surprising. But don't forget that this more complicated integral might have

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*.

This 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 $a_1, \ldots, a_n$. Then the points $a_1, \ldots a_n$ 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 12.2.9 on the page 868), 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 poynomials.

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, $i = 1, \ldots, k$. By a sequential use of the latter procedure with coprime polynomials $p(x)$ and $q(x)$, we obtain a representation of $f(x)/g(x)$ as a sum of fractions of the form

$$\frac{f(x)}{g(x)} = \frac{r_1(x)}{(x - a_1)^{n_1}} + \cdots + \frac{r_k(x)}{(x - a_k)^{m_j}},$$

where the degrees of the polynomials $r_i(x)$ are strictly smaller than the degrees of the denominators. Finally, each can be represented as a sum

$$\frac{r(x)}{(x - a)^n} = \frac{A_1}{x - a} + \frac{A_2}{(x - a)^2} + \cdots + \frac{A_n}{(x - a)^n}.$$

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 $A_1, A_2, \ldots$ 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 12.2.8 on page 866). The non-real

such a form that we may just be able to compute it. We want to determine the primitive function of function $f(x) = \operatorname{tg}^4 x$. Thus it's sensible to consider the substitution $u = \operatorname{tg} x$. We obtain

$$\int \frac{\sin^4 x}{\cos^4 x}\, dx = \begin{vmatrix} x = \operatorname{arctg} u \\ dx = \frac{du}{1+u^2} \end{vmatrix} = \int \frac{u^4}{1+u^2}\, du =$$
$$\int u^2 - 1 + \frac{1}{u^2+1}\, du =$$
$$\frac{u^3}{3} - u + \operatorname{arctg} u + C = \frac{\operatorname{tg}^3 x}{3} - \operatorname{tg} x + \operatorname{arctg}(\operatorname{tg} x) + C =$$
$$\frac{\operatorname{tg}^3 x}{3} - \operatorname{tg} x + x + C.$$

Case (d). We have

$$\int \frac{1 - \sqrt[3]{x} + \sqrt{x}}{\sqrt[6]{x^5} + x}\, dx = \begin{vmatrix} z^6 = x \\ 6z^5\, dz = dx \end{vmatrix} = \int \frac{1 - z^2 + z^3}{z^5 + z^6}\, 6z^5\, dz =$$
$$6 \int \frac{1 - z^2 + z^3}{1 + z}\, dz = 6 \int z^2 - 2z + 2 - \frac{1}{z+1}\, dz =$$
$$6 \left(\frac{z^3}{3} - z^2 + 2z - \ln|z+1| \right) + C =$$
$$2\sqrt{x} - 6\sqrt[3]{x} + 12\sqrt[6]{x} - 6\ln\left(\sqrt[6]{x} + 1\right) + C,$$

where we again easily determined by substitution (for $z \neq -1$)

$$\int \frac{dz}{z+1} = \begin{vmatrix} v = z + 1 \\ dv = dz \end{vmatrix} = \int \frac{dv}{v} = \ln|v| + C =$$
$$\ln|z+1| + C.$$

$\square$

6.B.11. By combining integration by parts and the substitution method, determine

(a) $\int x^3\, e^{-x^2}\, dx$, $x \in \mathbb{R}$;

(b) $\int x \arcsin x^2\, dx$, $x \in (-1, 1)$.

Solution. Case (a). The substitution method leads to the integral

$$\int x^3\, e^{-x^2}\, dx = \begin{vmatrix} t = -x^2 \\ dt = -2x\, dx \end{vmatrix} = \frac{1}{2} \int t\, e^t\, dt,$$

which can be easily computed by integrating by parts, yielding

$$\frac{1}{2} \int t\, e^t\, dt = \begin{vmatrix} F(t) = t & F'(t) = 1 \\ G'(t) = e^t & G(t) = e^t \end{vmatrix} =$$
$$\frac{1}{2} t\, e^t - \frac{1}{2} \int e^t\, dt =$$
$$\frac{1}{2} t\, e^t - \frac{1}{2} e^t + C = -\frac{1}{2} e^{-x^2} \left(x^2 + 1 \right) + C.$$

Case (b). Similarly, we obtain

$$\int x \arcsin x^2\, dx = \begin{vmatrix} t = x^2 \\ dt = 2x\, dx \end{vmatrix} = \frac{1}{2} \int \arcsin t\, dt =$$
$$\begin{vmatrix} F(t) = \arcsin t & F'(t) = \frac{1}{\sqrt{1-t^2}} \\ G'(t) = 1 & G(t) = t \end{vmatrix} =$$
$$\frac{1}{2} t \arcsin t - \frac{1}{2} \int \frac{t}{\sqrt{1-t^2}}\, dt = \begin{vmatrix} u = 1 - t^2 \\ du = -2t\, dt \end{vmatrix} =$$
$$\frac{1}{2} t \arcsin t + \frac{1}{4} \int \frac{du}{\sqrt{u}} = \frac{1}{2} t \arcsin t + \frac{1}{2} \sqrt{u} + C =$$
$$\frac{1}{2} t \arcsin t + \frac{1}{2} \sqrt{1 - t^2} + C = \frac{1}{2} x^2 \arcsin x^2 + \frac{1}{2} \sqrt{1 - x^4} + C.$$

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

$$\frac{A_1 x + B_1}{(x-a)^2 + b^2} + \cdots + \frac{A_n x + B_n}{((x-a)^2 + b^2)^n}$$

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 neighbourhood these functions are unbounded. We return to this problem later (see paragraph 6.2.16 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 ξ_i of the respective parts, i.e. $a = x_0 < x_1 < \cdots < x_n = b$ and $\xi_i \in [x_{i-1}, x_i]$, $i = 1, \ldots, n$. 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 $\Xi = (x_0, \ldots, x_n; \xi_1, \ldots, \xi_n)$ as

$$S_\Xi = \sum_{i=1}^{n} f(\xi_i) \cdot (x_i - x_{i-1}).$$

□

6.B.12. Compute the integral

$$\int \sqrt{1 - x^2} \, dx, \qquad x \in (-1, 1)$$

in two different ways.

Solution. Integration by parts yields

$$\int \sqrt{1 - x^2} \, dx = \begin{vmatrix} F(x) = \sqrt{1 - x^2} & F'(x) = \frac{-x}{\sqrt{1 - x^2}} \\ G'(x) = 1 & G(x) = x \end{vmatrix} =$$

$$x\sqrt{1 - x^2} + \int \frac{x^2}{\sqrt{1 - x^2}} \, dx = x\sqrt{1 - x^2} - \int \frac{1 - x^2 - 1}{\sqrt{1 - x^2}} \, dx =$$

$$x\sqrt{1 - x^2} - \int \sqrt{1 - x^2} \, dx + \int \frac{1}{\sqrt{1 - x^2}} \, dx =$$

$$x\sqrt{1 - x^2} - \int \sqrt{1 - x^2} \, dx + \arcsin x,$$

which implies

$$2 \int \sqrt{1 - x^2} \, dx = x\sqrt{1 - x^2} + \arcsin x + C,$$

i.e.

$$\int \sqrt{1 - x^2} \, dx = \tfrac{1}{2} \left(x\sqrt{1 - x^2} + \arcsin x \right) + C.$$

The substitution method along with (1) then yields

$$\int \sqrt{1 - x^2} \, dx = \begin{vmatrix} x = \sin y \\ dx = \cos y \, dy \end{vmatrix} = \int \sqrt{1 - \sin^2 y} \cdot$$

$$\cos y \, dy = \int \cos^2 y \, dy = \tfrac{1}{2} \left(y + \sin y \cdot \cos y \right) + C =$$

$$\tfrac{1}{2} \left(\sin y \cdot \sqrt{1 - \sin^2 y} + y \right) + C =$$

$$\tfrac{1}{2} \left(x\sqrt{1 - x^2} + \arcsin x \right) + C,$$

where $y \in (-\pi/2, \pi/2)$ for $x \in (-1, 1)$, thus among other things, we have

$$0 < \cos y = |\cos y| = \sqrt{\cos^2 y} = \sqrt{1 - \sin^2 y}.$$

□

6.B.13. Determine

$$\int e^{\sqrt{x}} \, dx, \qquad x > 0.$$

Solution. This problem can illustrate the possibilities of combining the substitution method and integration by parts (in the sscope of one problem. First we'll use the substitution $y = \sqrt{x}$ to get rid of the root from the argument of the exponential function. That leads to the integral

$$\int e^{\sqrt{x}} \, dx = \begin{vmatrix} y^2 = x \\ 2y \, dy = dx \end{vmatrix} = 2 \int y \, e^y \, dy.$$

Now by using integration by parts, we'll compute

$$\int y \, e^y \, dy = \begin{vmatrix} F(y) = y & F'(y) = 1 \\ G'(y) = e^y & G(y) = e^y \end{vmatrix} =$$

$$y \, e^y - \int e^y \, dy = y \, e^y - e^y + C.$$

Thus in total, we have

$$\int e^{\sqrt{x}} \, dx = 2y \, e^y - 2 \, e^y + C = 2 \, e^{\sqrt{x}} \left(\sqrt{x} - 1 \right) + C.$$
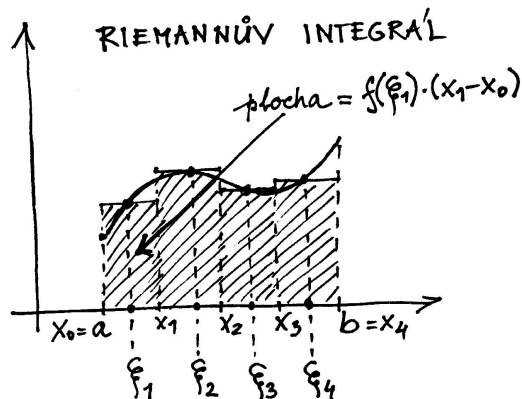
RIEMANN INTEGRAL[4]

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 δ_k approaching zero, the limit

$$\lim_{k \to \infty} S_{\Xi_k} = S$$

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:

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 $\int_a^b C \cdot f(x) \, dx$ also exists and

$$\int_a^b C \cdot f(x) \, dx = C \cdot \int_a^b f(x) \, dx.$$

[4]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.

6.B.14. Prove that

$$\frac{1}{2}\sin^4 x = -\frac{1}{4}\cos(2x) + \frac{1}{16}\cos(4x) + \frac{3}{16}.$$

Solution. Easier than to compare the given expressions directly is to show that the functions on the right and left hand side have the same derivatives. We have $L' = 2\cos x \sin^3 x = \sin(2x)\sin^2 x$,

$$P' = \frac{1}{2}\sin(2x) + \frac{1}{4}\sin(4x) = \sin 2x(\frac{1}{2} + \frac{1}{2}\cos(2x)) = \sin(2x)\sin^2 x.$$ Hence the left and the right hand side differ by a constant. This constant can be determined by comparing the values at one point, for example 0. Both functions are zero at zero, thus they are equal. □

Integration of rational functions.

The key to integration of rational functions lies in decomposition of a rational function as a sum of a simple rational functions, which we know, how to integrate. Let us decompose some rational functions:

6.B.15. Carry out the suggested division of polynomials

$$\frac{2x^5 - x^4 + 3x^2 - x + 1}{x^2 - 2x + 4}$$

for $x \in \mathbb{R}$. ○

6.B.16. Express the function

$$y = \frac{3x^4 + 2x^3 - x^2 + 1}{3x + 2}$$

as a sum of a polynomial and a rational function. ○

6.B.17. Decompose the rational expression

(a) $\frac{4x^2 + 13x - 2}{x^3 + 3x^2 - 4x - 12}$;

(b) $\frac{2x^5 + 5x^3 - x^2 + 2x - 1}{x^6 + 2x^4 + x^2}$

into partial fractions. ○

6.B.18. Express the function

$$y = \frac{2x^3 + 6x^2 + 3x - 6}{x^4 - 2x^3}$$

in the form of partial fractions. ○

6.B.19. Decompose the expression

$$\frac{7x^2 - 10x + 37}{x^3 - 3x^2 + 9x + 13}$$

into partial fractions. ○

6.B.20. Express the rational function

$$y = \frac{-5x + 2}{x^4 - x^3 + 2x^2}$$

in the form of a sum of partial fractions. ○

6.B.21. Decompose the function

$$y = \frac{1}{x^3(x+1)}$$

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 sums is multiplied by the constant C. So the claim follows from the elementary properties of limits. □

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

into partial fractions.

6.B.22. Determine the form of the decomposition of the rational function
$$y = \frac{2x^2 - 114}{(x-2)\,x^2\,(3x^2+x+4)^2}$$
into partial fractions. Don't compute the undetermined coefficients!

6.B.23. Express the function
$$y = \frac{x^4 + 6x^2 + x - 2}{x^4 - 2x^3}$$
as a sum of a polynomial and a proper rational function Q. Then express the obtained function Q in the form of a sum of partial fractions.

6.B.24. Write the primitive function to the rational function

(a) $y = \frac{3}{x-2}$, $\quad x \neq 2$;

(b) $y = -\frac{2}{(x-2)^3}$, $\quad x \neq 2$.

6.B.25. Integrate

(a) $\int \frac{6}{x-2}\,dx$, $x \neq 2$;

(b) $\int \frac{6}{(x+4)^3}\,dx$, $x \neq -4$;

(c) $\int \frac{3x+7}{x^2-4x+15}\,dx$, $x \in \mathbb{R}$;

(d) $\int \frac{30x-77}{(x^2-6x+13)^2}\,dx$, $x \in \mathbb{R}$.

Solution. Cases (a), (b). We have
$$\int \frac{6}{x-2}\,dx = \left| \begin{array}{l} y = x-2 \\ dy = dx \end{array} \right| = \int \frac{6}{y}\,dy = 6\ln|y| + C = $$
$$6\ln|x-2| + C$$
and similarly
$$\int \frac{6}{(x+4)^3}\,dx = \left| \begin{array}{l} y = x+4 \\ dy = dx \end{array} \right| = \int \frac{6}{y^3}\,dy = \frac{6}{-2y^2} + C = $$
$$-\frac{3}{(x+4)^2} + C.$$

We can see that integrating the partial fractions which correspond to real roots of a denominator of rational function is very easy. Moreover, without loss of generality we can obtain
$$\int \frac{A}{x-x_0}\,dx = \left| \begin{array}{l} y = x - x_0 \\ dy = dx \end{array} \right| = \int \frac{A}{y}\,dy = A\ln|y| + C = $$
$$A\ln|x - x_0| + C$$
and
$$\int \frac{A}{(x-x_0)^n}\,dx = \left| \begin{array}{l} y = x - x_0 \\ dy = dx \end{array} \right| = \int \frac{A}{y^n}\,dy = $$
$$\frac{Ay^{-n+1}}{-n+1} + C = \frac{A}{(1-n)(x-x_0)^{n-1}} + C$$
for all $A, x_0 \in \mathbb{R}$, $n \geq 2$, $n \in \mathbb{N}$.

Case (c). Now we are to integrate a partial fraction corresponding to a pair of complex conjugate roots. Thus in the

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 $\Xi = (x_0, \ldots, x_n)$ is given by the expression
$$S_{\Xi,\sup} = \sum_{i=1}^{n} \Big(\sup_{x_{i-1} \leq \xi \leq x_i} f(\xi) \Big)(x_i - x_{i-1})$$
$$= \sum_{i=1}^{n} M_i(x_i - x_{i-1}).$$

The *lower Riemann sum* is
$$S_{\Xi,\inf} = \sum_{i=1}^{n} \Big(\inf_{x_{i-1} \leq \xi \leq x_i} f(\xi) \Big)(x_i - x_{i-1})$$
$$= \sum_{i=1}^{n} m_i(x_i - x_{i-1}).$$

For each partition $\Xi = (x_0, \ldots, x_n; \xi_1, \ldots, \xi_n)$ with representatives, there are the inequalities

(1) $$S_{\Xi,\inf} \leq S_{\Xi,\xi} \leq S_{\Xi,\sup}$$

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 exists and they will be equal. This is indeed true for all bounded functions:

Theorem. *Let the function f be bounded on a closed interval $[a, b]$. Then*
$$S_{\sup} = \inf_{\Xi} S_{\Xi,\sup}, \qquad S_{\inf} = \sup_{\Xi} S_{\Xi,\inf}$$
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 $S_{\sup} = S_{\inf}$.

PROOF. First, notice that $S_{\sup}$ is well defined since it is the infimum of a set of real values bounded from below by any of the $S_{\Xi,\inf}$. Similarly for the value $S_{\inf}$, which is bounded from above by any of $S_{\Xi,\sup}$.

Refine a partition Ξ_1 to Ξ_2 by adding new points. Then
$$S_{\Xi_1,\sup} \geq S_{\Xi_2,\sup}, \qquad S_{\Xi_1,\inf} \leq S_{\Xi_2,\inf}.$$

By the definition of the infimum, there are sequences of partitions Ξ_k for which $S_{\sup}$ is the limit of the sums $S_{\Xi_k,\sup}$. Moreover, every two partitions have a common refinement. Thus it may be assumed that Ξ_k in the sequence is always a partition obtained by refining the previous one. Hence the sums $S_{\Xi_k,\sup}$ form a non-increasing sequence of real numbers converging to $S_{\sup}$.

denominator there is a polynomial of degree 2 and in the numerator at most 1. If it's of degree 1, we'll write the partial fraction so that we'll have a multiple of the derivative of the denominator in the numerator and add to it the fraction, in whose numerator there is only a constant. This way we'll obtain

$$\int \frac{3x+7}{x^2-4x+15}\, dx = \frac{3}{2} \int \frac{2x-4}{x^2-4x+15}\, dx + 13 \int \frac{dx}{x^2-4x+15} =$$
$$\frac{3}{2} \ln \left(x^2 - 4x + 15 \right) + 13 \int \frac{dx}{(x-2)^2+11} =$$

$$\frac{3}{2} \ln \left(x^2 - 4x + 15 \right) + \frac{13}{11} \int \frac{dx}{\left(\frac{x-2}{\sqrt{11}}\right)^2+1} = \left| \begin{array}{c} y = \frac{x-2}{\sqrt{11}} \\ dy = \frac{dx}{\sqrt{11}} \end{array} \right| =$$

$$\frac{3}{2} \ln \left(x^2 - 4x + 15 \right) + \frac{13}{\sqrt{11}} \int \frac{dy}{y^2+1} = \frac{3}{2} \ln \left(x^2 - 4x + 15 \right) +$$
$$\frac{13}{\sqrt{11}} \operatorname{arctg} y + C = \frac{3}{2} \ln \left(x^2 - 4x + 15 \right) + \frac{13}{\sqrt{11}} \operatorname{arctg} \frac{x-2}{\sqrt{11}} + C.$$

Again, we can generally express

$$\int \frac{Ax+B}{(x-x_0)^2+a^2}\, dx =$$
$$\frac{A}{2} \int \frac{2(x-x_0)}{(x-x_0)^2+a^2}\, dx + (B + Ax_0) \int \frac{1}{(x-x_0)^2+a^2}\, dx$$

and compute

$$\int \frac{2(x-x_0)}{(x-x_0)^2+a^2}\, dx = \left| \begin{array}{c} y = (x-x_0)^2 + a^2 \\ dy = 2(x-x_0)\, dx \end{array} \right| = \int \frac{dy}{y} =$$
$$\ln |y| + C = \ln[(x-x_0)^2 + a^2] + C,$$

$$\int \frac{1}{(x-x_0)^2+a^2}\, dx = \frac{1}{a^2} \int \frac{dx}{\left(\frac{x-x_0}{a}\right)^2+1} = \left| \begin{array}{c} z = \frac{x-x_0}{a} \\ dz = \frac{dx}{a} \end{array} \right| =$$
$$\frac{1}{a} \int \frac{dz}{z^2+1} = \frac{1}{a} \operatorname{arctg} z + C = \frac{1}{a} \operatorname{arctg} \frac{x-x_0}{a} + C,$$

i.e.

$$\int \frac{Ax+B}{(x-x_0)^2+a^2}\, dx =$$
$$\frac{A}{2} \ln \left((x-x_0)^2 + a^2 \right) + \frac{B+Ax_0}{a} \operatorname{arctg} \frac{x-x_0}{a} + C,$$

where the values $A, B, x_0 \in \mathbb{R}, a > 0$ are arbitrary.

Case (d). All that is left are the partial fractions for multiple complex roots in the form of

$$\frac{Ax+B}{[(x-x_0)^2+a^2]^n}, \qquad A, B, x_0 \in \mathbb{R},\ a > 0, n \in \mathbb{N} \setminus \{1\},$$

which can be analogically simplified to

$$\frac{A}{2} \cdot \frac{2(x-x_0)}{[(x-x_0)^2+a^2]^n} + (B+Ax_0) \cdot \frac{1}{[(x-x_0)^2+a^2]^n}.$$

Then we'll determine

$$\int \frac{2(x-x_0)}{[(x-x_0)^2+a^2]^n}\, dx = \left| \begin{array}{c} y = (x-x_0)^2 + a^2 \\ dy = 2(x-x_0)\, dx \end{array} \right| = \int \frac{dy}{y^n} =$$
$$\frac{1}{(1-n)y^{n-1}} + C = \frac{1}{(1-n)[(x-x_0)^2+a^2]^{n-1}} + C$$

and

$$K_n(x_0, a) := \int \frac{1}{[(x-x_0)^2+a^2]^n}\, dx =$$
$$\left| \begin{array}{cc} F(x) = \frac{1}{[(x-x_0)^2+a^2]^n} & F'(x) = \frac{-2n(x-x_0)}{[(x-x_0)^2+a^2]^{n+1}} \\ G'(x) = 1 & G(x) = x - x_0 \end{array} \right| =$$
$$\frac{x-x_0}{[(x-x_0)^2+a^2]^n} + 2n \int \frac{(x-x_0)^2+a^2}{[(x-x_0)^2+a^2]^{n+1}} -$$
$$\frac{a^2}{[(x-x_0)^2+a^2]^{n+1}}\, dx =$$
$$\frac{x-x_0}{[(x-x_0)^2+a^2]^n} + 2n \left(K_n(x_0, a) - a^2 K_{n+1}(x_0, a) \right),$$

A similar argument applies to $S_{\inf}$. Hence the values

$$S_{\sup} = \inf_{\Xi} S_{\Xi,\sup}, \qquad S_{\inf} = \sup_{\Xi} S_{\Xi,\inf}$$

are good candidates for the limits of upper and lower sums.

Next, consider a fixed partition Ξ with n inner partitioning points of the interval $[a, b]$, and another partition Ξ_1, whose norm is a small number δ. In the common refinement Ξ_2, there will be only n intervals contributing to the sum $S_{\Xi_2,\sup}$ by eventually smaller contribution than in the case of Ξ_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 δ of the partition. Hence when choosing δ sufficiently small, the distance of $S_{\Xi_1,\sup}$ from $S_{\sup}$ will not be larger than twice the distance of $S_{\Xi,\sup}$ from $S_{\sup}$.

Finally, return to the sequence of partitions Ξ_k as chosen above, and choose an $\varepsilon > 0$. Then there is some $m \in \mathbb{N}$ such that the distance of $S_{\Xi_k,\sup}$ from $S_{\sup}$ is less than ε for all $k \geq m$. Hence for arbitrary partition Ξ with appropriately small norm $\delta > 0$ the distance of $S_{\Xi,\sup}$ from $S_{\sup}$ does not exceed 2ε.

In summary, for arbitrary $2\varepsilon > 0$, there is $\delta > 0$ such that for all partitions with norm at most δ the inequality $|S_{\Xi,\sup} - S_{\sup}| < 2\varepsilon$ holds. This is exactly the statement that the number $S_{\sup}$ 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 $S_{\sup} = S_{\inf}$, then all Riemann sums of 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 Ξ_k and $\bar{\bar{\Xi}}_k$ 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 $S_{\sup} > S_{\inf}$. □

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 $S_{\sup} = S_{\inf}$.

From the definition of continuity, for every fixed 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:

which implies

$$K_{n+1}(x_0, a) =$$
$$\frac{1}{a^2}\left(\frac{2n-1}{2n} K_n(x_0, a) + \frac{1}{2n}\frac{x-x_0}{[(x-x_0)^2+a^2]^n}\right),$$

which clearly also holds for $n = 1$. The last recurrent formula can be extended with the integral (derived in case (c))

$$K_1(x_0, a) = \frac{1}{a}\operatorname{arctg}\frac{x-x_0}{a} + C.$$

In the given problem we have

$$\int \frac{30x-77}{(x^2-6x+13)^2}\, dx =$$
$$15 \int \frac{2x-6}{(x^2-6x+13)^2}\, dx + 13 \int \frac{1}{(x^2-6x+13)^2}\, dx$$

and furthermore

$$\int \frac{2x-6}{(x^2-6x+13)^2}\, dx = \left| \begin{array}{l} y = x^2 - 6x + 13 \\ dy = (2x-6)\, dx \end{array} \right| = \int \frac{dy}{y^2} =$$
$$-\frac{1}{y} + C = -\frac{1}{x^2-6x+13} + C,$$

$$\int \frac{1}{(x^2-6x+13)^2}\, dx = \int \frac{dx}{[(x-3)^2+2^2]^2} =$$
$$\frac{1}{2^2}\left(\frac{2-1}{2} K_1(3,2) + \frac{1}{2}\frac{x-3}{(x-3)^2+2^2}\right) =$$
$$\frac{1}{4}\left(\frac{1}{4}\operatorname{arctg}\frac{x-3}{2} + C + \frac{1}{2}\frac{x-3}{x^2-6x+13}\right) =$$
$$\frac{1}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{1}{8}\frac{x-3}{x^2-6x+13} + C.$$

In total, we have

$$\int \frac{30x-77}{(x^2-6x+13)^2}\, dx = -\frac{15}{x^2-6x+13} + \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} +$$
$$\frac{13}{8}\frac{x-3}{x^2-6x+13} + C = \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{13x-159}{8(x^2-6x+13)} + C.$$

$\square$

6.B.26. Integrate the rational functions

(a) $\int \frac{x^3+1}{x(x-1)^3}\, dx$, $x \neq 0$, $x \neq 1$;

(b) $\int \frac{x-4}{5x^2+6x+3}\, dx$, $x \in \mathbb{R}$;

(c) $\int \frac{1}{(x-4)(x-2)(x^2+2x+2)}\, dx$, $x \neq 2$, $x \neq 4$;

(d) $\int \frac{x}{x^4-x^3-x+1}\, dx$, $x \neq 1$;

(e) $\int \frac{2x+1}{(x^2+4x+13)^2}\, dx$, $x \in \mathbb{R}$;

(f) $\int \frac{5x^2-12}{x^4-12x^3+62x^2-156x+169}\, dx$, $x \in \mathbb{R}$.

Solution. We'll compute all the given integrals in the way we can always use when integrating rational functions. We won't use any specific simplification or substitution. Even the recurrent formula for $K_{n+1}(x_0, a)$, which we derived in a general form, will be used only for $x_0 = 0$, $a = 1$ (and also when $n = 0$). Using the aforementioned procedures, we obtain

(a)

$$\int \frac{x^3+1}{x(x-1)^3}\, dx = 2 \int \frac{dx}{x-1} + \int \frac{dx}{(x-1)^2} + 2 \int \frac{dx}{(x-1)^3} -$$
$$\int \frac{dx}{x} = 2\ln|x-1| - \frac{1}{x-1} - \frac{1}{(x-1)^2} - \ln|x| + C;$$

(b)

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.

STEJNOMĚRNÁ SPOJITOST
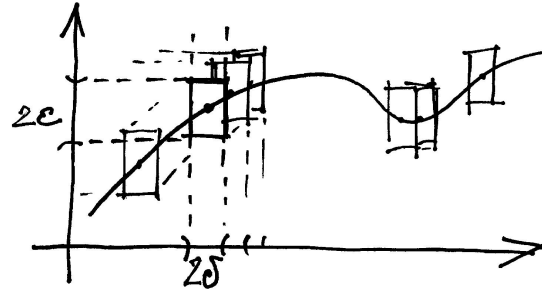
Theorem. *Each continuous function on a finite closed interval $[a, b]$ is uniformly continuous.*

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 $x_1, \ldots, x_k$. Choose δ 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 ε and δ as in the definition of the uniform continuity of f. Consider any partition Ξ with n intervals and norm at most δ. Then, writing $J_i = [x_{i-1}, x_i]$,

$$\left| \sum_{i=1}^{n} \sup_{\xi \in J_i} f(\xi)(x_i - x_{i-1}) - \sum_{i=1}^{n} \inf_{\xi \in J_i} f(\xi)(x_i - x_{i-1}) \right|$$
$$= \sum_{i=1}^{n}\left(\sup_{\xi \in J_i} f(\xi) - \inf_{\xi \in J_i} f(\xi)\right)(x_i - x_{i-1})$$
$$\leq \varepsilon \cdot (b - a).$$

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

$$\int \frac{x-4}{5x^2+6x+3}\, dx =$$

$$\frac{1}{10}\int \frac{10x+6}{5x^2+6x+3}\, dx - \frac{23}{5}\int \frac{dx}{5x^2+6x+3} =$$

$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23}{25}\int \frac{dx}{\left(x+\frac{3}{5}\right)^2+\frac{6}{25}} =$$

$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23}{6}\int \frac{dx}{\left(\frac{5x+3}{\sqrt6}\right)^2+1} =$$

$$\left| \begin{array}{c} t=\frac{5x+3}{\sqrt6} \\ dt=\frac{5}{\sqrt6}\, dx \end{array}\right| =$$

$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt6}{30}\int \frac{dt}{t^2+1} =$$

$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt6}{30}\,\operatorname{arctg} t + C =$$

$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt6}{30}\,\operatorname{arctg}\frac{5x+3}{\sqrt6} + C;$$

(c)

$$\int \frac{dx}{(x-4)(x-2)(x^2+2x+2)} = \frac{1}{52}\int\frac{dx}{x-4} - \frac{1}{20}\int\frac{dx}{x-2} +$$

$$\frac{1}{130}\int\frac{4x+11}{x^2+2x+2}\,dx = \frac{1}{52}\ln\,|\,x-4\,| - \frac{1}{20}\ln\,|\,x-2\,| +$$

$$\frac{1}{130}\left(2\int\frac{2x+2}{x^2+2x+2}\,dx + 7\int\frac{dx}{x^2+2x+2}\right) =$$

$$\frac{1}{260}\ln\left|\frac{(x-4)^5}{(x-2)^{13}}\right| + \frac{2}{130}\ln\left(x^2+2x+2\right) +$$

$$\frac{7}{130}\int\frac{dx}{(x+1)^2+1} = \left|\begin{array}{c} t=x+1 \\ dt=dx\end{array}\right| =$$

$$\frac{1}{260}\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + \frac{7}{130}\int\frac{dt}{t^2+1} =$$

$$\frac{1}{260}\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + \frac{7}{130}\,\operatorname{arctg} t + C =$$

$$\frac{1}{260}\left[\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + 14\,\operatorname{arctg}\left(x+1\right)\right] + C;$$

(d)

$$\int \frac{x}{x^4-x^3-x+1}\,dx = \frac{1}{3}\int\frac{dx}{(x-1)^2} - \frac{1}{3}\int\frac{dx}{x^2+x+1} =$$

$$-\frac{1}{3(x-1)} - \frac{1}{3}\int\frac{dx}{\left(x+\frac{1}{2}\right)^2+\frac{3}{4}} =$$

$$-\frac{1}{3(x-1)} - \frac{4}{9}\int\frac{dx}{\left(\frac{2x+1}{\sqrt3}\right)^2+1} = \left|\begin{array}{c} t=\frac{2x+1}{\sqrt3} \\ dt=\frac{2}{\sqrt3}\, dx\end{array}\right| =$$

$$-\frac{1}{3(x-1)} - \frac{2}{3\sqrt3}\int\frac{dt}{t^2+1} = -\frac{1}{3(x-1)} - \frac{2}{3\sqrt3}\,\operatorname{arctg} t + C =$$

$$-\frac{1}{3(x-1)} - \frac{2}{3\sqrt3}\,\operatorname{arctg}\frac{2x+1}{\sqrt3} + C;$$

(e)

$$\int \frac{2x+1}{(x^2+4x+13)^2}\,dx = \int\frac{2x+4}{(x^2+4x+13)^2}\,dx -$$

$$3\int\frac{dx}{(x^2+4x+13)^2} = \left|\begin{array}{c} t=x^2+4x+13 \\ dt=(2x+4)\, dx\end{array}\right| =$$

$$\int\frac{dt}{t^2} - 3\int\frac{dx}{\left[(x+2)^2+9\right]^2} = -\frac{1}{t} - \frac{1}{27}\int\frac{dx}{\left[\left(\frac{x+2}{3}\right)^2+1\right]^2} =$$

$$\left|\begin{array}{c} u=\frac{x+2}{3} \\ du=\frac{1}{3}\, dx\end{array}\right| = -\frac{1}{x^2+4x+13} - \frac{1}{9}\int\frac{du}{(u^2+1)^2} =$$

$$-\frac{1}{x^2+4x+13} - \frac{1}{9}\left(\frac{1}{2}\operatorname{arctg} u + \frac{1}{2}\,\frac{u}{u^2+1}\right) + C =$$

$$-\frac{1}{x^2+4x+13} - \frac{1}{18}\operatorname{arctg}\frac{x+2}{3} - \frac{1}{18}\,\frac{\frac{x+2}{3}}{\left(\frac{x+2}{3}\right)^2+1} + C =$$

$$-\frac{1}{18}\operatorname{arctg}\frac{x+2}{3} - \frac{1}{6}\,\frac{x+8}{x^2+4x+13} + C;$$

(f)

$$\int \frac{5x^2-12}{x^4-12x^3+62x^2-156x+169}\,dx = \int\frac{5x^2-12}{(x^2-6x+13)^2}\,dx =$$

$$5\int\frac{dx}{x^2-6x+13} + \int\frac{30x-77}{(x^2-6x+13)^2}\,dx = 5\int\frac{dx}{(x-3)^2+4} +$$

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 $0 \le \Delta x < \delta$ 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 Δx 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 $\xi \in [x, x+\Delta]$ are distant from $f(x)$ by at most ε. Hence the Riemann integral in question cannot be different from $f(x)\Delta$ by more then $\varepsilon\Delta$. 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 $[x-\Delta, x]$. This finishes the proof of the theorem 6.2.9.

6.2.13. Important remarks. (1) Theorems 6.2.9 and 6.2.8 claim that the Riemann integral is a linear map

$$\int : C[a,b] \to \mathbb{R}$$

from the vector space of continuous functions on the interval $[a,b]$ to real numbers. Hence it is a linear form on the 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 $(0,1)$.

$15 \int \frac{2x-6}{(x^2-6x+13)^2} \, dx + 13 \int \frac{dx}{(x^2-6x+13)^2} =$
$\frac{5}{4} \int \frac{dx}{\left(\frac{x-3}{2}\right)^2+1} + 15 \int \frac{2x-6}{(x^2-6x+13)^2} \, dx +$
$13 \int \frac{dx}{\left[(x-3)^2+4\right]^2} =$

$\left| \begin{array}{c|c} t = \frac{x-3}{2} & u = x^2 - 6x + 13 \\ dt = \frac{1}{2} dx & du = (2x-6) \, dx \end{array} \right| =$

$\frac{5}{2} \int \frac{dt}{t^2+1} + 15 \int \frac{du}{u^2} + \frac{13}{16} \int \frac{dx}{\left[\left(\frac{x-3}{2}\right)^2+1\right]^2} =$

$\frac{5}{2} \operatorname{arctg} t - \frac{15}{u} + \frac{13}{8} \int \frac{dt}{[t^2+1]^2} = \frac{5}{2} \operatorname{arctg} \frac{x-3}{2} -$
$\frac{15}{x^2-6x+13} + \frac{13}{8} \left(\frac{1}{2} \operatorname{arctg} t + \frac{1}{2} \frac{t}{t^2+1} \right) + C =$
$\frac{5}{2} \operatorname{arctg} \frac{x-3}{2} - \frac{15}{x^2-6x+13} + \frac{13}{16} \operatorname{arctg} \frac{x-3}{2} + \frac{13}{16} \frac{\frac{x-3}{2}}{\left(\frac{x-3}{2}\right)^2+1} +$

$C = \frac{5}{2} \operatorname{arctg} \frac{x-3}{2} + \frac{13}{16} \operatorname{arctg} \frac{x-3}{2} - \frac{15}{x^2-6x+13} +$
$\frac{13}{8} \frac{x-3}{(x-3)^2+4} + C = \frac{53}{16} \operatorname{arctg} \frac{x-3}{2} + \frac{13x-159}{8(x^2-6x+13)} + C.$

$\square$

6.B.27. Compute

$$\int \frac{x}{(x-1)^2(x^2+2x+2)} \, dx, \qquad x \neq 1.$$

Solution. Because the degree of the polynomial in the denominator is lower than in the numerator, these polynomials don't have a common root and we know the representation of the denominator in the form of a product of root factors, we know the form of the decomposition of the integrated function into partials fractions

$$\frac{x}{(x-1)^2(x^2+2x+2)} = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{Cx+D}{x^2+2x+2}$$

for $A, B, C, D \in \mathbb{R}$. If we multiply this equation by the denominator of the left hand side, we'll obtain the identity

$$x = A(x-1)\left(x^2+2x+2\right) + B\left(x^2+2x+2\right) + (Cx+D)(x-1)^2,$$

which hold for all $x \in \mathbb{R} \smallsetminus \{1\}$. But on its both sided there are polynomials, so the equality must also occur for $x = 1$. By substituing this value we immediately get $1 = B(1+2+2)$, i.e. $B = 1/5$.

We could choose other real (eventually complex) numbers and substitute them into the given equation, but we cannot expect to directly determine another of the variables (if we don't substitute a root of the denominator). Thus we'll rather compare the coefficients at the same powers of the polynomials

$$x - \frac{1}{5}\left(x^2+2x+2\right) = -\frac{1}{5}x^2 + \frac{3}{5}x - \frac{2}{5},$$

$$A(x-1)\left(x^2+2x+2\right) + (Cx+D)(x-1)^2 =$$
$$(A+C)x^3 + (A-2C+D)x^2 + (C-2D)x - 2A + D,$$

(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

$$\int_a^b f(x) \, dx = F(b) - F(a) = F'(c)(b-a) = f(c)(b-a).$$

This statement can be derived directly from the definition of the Riemann integral. It can be used in the final step of the proof of the fundamental theorem of integral calculus.

6.2.14. 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 equation with *separated variables*

$$y' = f(x)g(y)$$

which leads to a system of equations

$$
\begin{array}{rcrcrcr}
0 &=& A &+& C, & & \\
-1/5 &=& A &-& 2C &+& D, \\
3/5 &=& & & C &-& 2D, \\
-2/5 &=& -2A & & &+& D.
\end{array}
$$

Note that this system must have exactly one solution (which is uniquely determined by any three of the given equations). The sought solution is then

$$
A = \tfrac{1}{25}, \qquad C = -\tfrac{1}{25}, \qquad D = -\tfrac{8}{25}.
$$

Thus

$$
\int \frac{x}{(x-1)^2(x^2+2x+2)} \, dx =
$$
$$
\int \frac{dx}{25(x-1)} + \int \frac{dx}{5(x-1)^2} - \int \frac{x+8}{25(x^2+2x+2)} \, dx = \tfrac{1}{25}\ln|x-1| - \tfrac{1}{5(x-1)} - \tfrac{1}{50}\ln\left(x^2+2x+2\right) - \tfrac{7}{25}\operatorname{arctg}(x+1) + C,
$$

where we used

$$
\int \frac{x+8}{x^2+2x+2} \, dx = \int \frac{\tfrac{1}{2}(2x+2)}{x^2+2x+2} + \frac{7}{x^2+2x+2} \, dx =
$$
$$
\tfrac{1}{2}\int \frac{2x+2}{x^2+2x+2}\, dx +
$$
$$
7\int \frac{1}{(x+1)^2+1}\, dx = \tfrac{1}{2}\ln\left(x^2+2x+2\right) + 7\operatorname{arctg}(x+1) + C.
$$

$\square$

6.B.28. Determine

(a) $\int \frac{x^3+2x^2+x-1}{x^2-x+1} \, dx$, $x \in \mathbb{R}$;

(b) $\int \frac{x^8}{x^8-1} \, dx$, $x \neq \pm 1$.

Solution. Case (a). First we must do the division of polynomials

$$
(x^3 + 2x^2 + x - 1) : (x^2 - x + 1) = x + 3 + \tfrac{3x-4}{x^2-x+1},
$$

to consider a proper rational function (with the degree of the numerator lower then the degree of the denominator). Now we'll compute

$$
\int \frac{x^3+2x^2+x-1}{x^2-x+1} \, dx = \int x + 3 \, dx + \int \frac{3x-4}{x^2-x+1} \, dx =
$$
$$
\tfrac{x^2}{2} + 3x + \tfrac{3}{2}\int \frac{2x-1}{x^2-x+1} \, dx - \tfrac{5}{2}\int \frac{dx}{\left(x-\tfrac{1}{2}\right)^2 + \left(\tfrac{\sqrt{3}}{2}\right)^2} =
$$
$$
\tfrac{x^2}{2} + 3x + \tfrac{3}{2}\ln\left(x^2-x+1\right) - \tfrac{5}{\sqrt{3}}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} + C.
$$

Case (b). We have

$$
\int \frac{x^8}{x^8-1} \, dx = \int 1 \, dx + \tfrac{1}{8}\int \frac{dx}{x-1} - \tfrac{1}{8}\int \frac{dx}{x+1} - \tfrac{1}{4}\int \frac{dx}{x^2+1} +
$$
$$
\tfrac{1}{8}\int \frac{\sqrt{2}x-2}{x^2-\sqrt{2}x+1} \, dx - \tfrac{1}{8}\int \frac{\sqrt{2}x+2}{x^2+\sqrt{2}x+1} \, dx = x + \tfrac{1}{8}\ln|x-1| -
$$
$$
\tfrac{1}{8}\ln|x+1| - \tfrac{1}{4}\operatorname{arctg} x + \tfrac{\sqrt{2}}{16}\int \frac{2x-\sqrt{2}}{x^2-\sqrt{2}x+1} \, dx -
$$
$$
\tfrac{1}{8}\int \frac{dx}{\left(x-\tfrac{\sqrt{2}}{2}\right)^2 + \left(\tfrac{\sqrt{2}}{2}\right)^2} - \tfrac{\sqrt{2}}{16}\int \frac{2x+\sqrt{2}}{x^2+\sqrt{2}x+1} \, dx -
$$
$$
\tfrac{1}{8}\int \frac{dx}{\left(x+\tfrac{\sqrt{2}}{2}\right)^2 + \left(\tfrac{\sqrt{2}}{2}\right)^2} = x + \tfrac{1}{8}\ln|x-1| - \tfrac{1}{8}\ln|x+1| -
$$
$$
\tfrac{1}{4}\operatorname{arctg} x + \tfrac{\sqrt{2}}{16}\ln\left(x^2-\sqrt{2}x+1\right) - \tfrac{\sqrt{2}}{8}\operatorname{arctg}\left(\sqrt{2}x-1\right) -
$$
$$
\tfrac{\sqrt{2}}{16}\ln\left(x^2+\sqrt{2}x+1\right) - \tfrac{\sqrt{2}}{8}\operatorname{arctg}\left(\sqrt{2}x+1\right) + C.
$$

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.

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 $y(x) = D\,\mathrm{e}^x$ 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 $y(x) = \mathrm{e}^x$.

6.2.15. Analytic solutions. In the next part of this chapter, we shall prove that 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 $f(x) = \sum_{n=0}^{\infty} \frac{a_n}{n!} x^n$ is

$$
y(x) = \sum_{n=0}^{\infty} \frac{1}{n+1} a_n x^{n+1} + y(0),
$$

where $y(0)$ is the free integration constant. The solution is defined on the coverengce domain of the power series. 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

$$
(1) \qquad\qquad y' = f(y)
$$

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

$$
y'(x_0) = f(y_0)
$$
$$
y''(x_0) = f'(y)y'|_{x=x_0} = f'(y)f(y)|_{x=x_0} = f'(y_0)f(y_0)
$$
$$
y'''(x_0) = \left(f''(y)y'f(y) + f'(y)f'(y)y'\right)|_{x=x_0}
$$
$$
= f''(y_0)(f(y_0))^2 + (f'(y_0))^2 f(y_0)
$$
$$
\vdots
$$

6.B.29. Compute

$$\int \frac{2x^4+2x^2-5x+1}{x(x^2-x+1)^2} \, dx, \qquad x \neq 0.$$

Solution. We have

$$\int \frac{2x^4+2x^2-5x+1}{x(x^2-x+1)^2} \, dx =$$

$$\int \frac{dx}{x} + \int \frac{x+3}{x^2-x+1} \, dx + \int \frac{x-6}{(x^2-x+1)^2} \, dx =$$

$$\ln|x| + \frac{1}{2}\int \frac{2x-1}{x^2-x+1}\,dx + \frac{7}{2}\int \frac{dx}{x^2-x+1} + \frac{1}{2}\int \frac{2x-1}{(x^2-x+1)^2}\,dx -$$

$$\frac{11}{2}\int \frac{dx}{(x^2-x+1)^2} = \left| \begin{array}{l} t = x^2-x+1 \\ dt = (2x-1)\,dx \end{array} \right| =$$

$$\ln|x| + \frac{1}{2}\ln\left(x^2-x+1\right) + \frac{7}{2}\int \frac{dx}{(x-\frac{1}{2})^2+\frac{3}{4}} + \frac{1}{2}\int \frac{dt}{t^2} -$$

$$\frac{11}{2}\int \frac{dx}{\left[(x-\frac{1}{2})^2+\frac{3}{4}\right]^2} = \ln\left| x\sqrt{x^2-x+1}\right| +$$

$$\frac{14}{3}\int \frac{dx}{\left(\frac{2x-1}{\sqrt{3}}\right)^2+1} - \frac{1}{2t} - \frac{88}{9}\int \frac{dx}{\left[\left(\frac{2x-1}{\sqrt{3}}\right)^2+1\right]^2} =$$

$$\left| \begin{array}{l} u = \frac{2x-1}{\sqrt{3}} \\ du = \frac{2}{\sqrt{3}}\,dx \end{array} \right| = \ln\left| x\sqrt{x^2-x+1}\right| + \frac{7\sqrt{3}}{3}\int \frac{du}{u^2+1} -$$

$$\frac{1}{2(x^2-x+1)} - \frac{44\sqrt{3}}{9}\int \frac{du}{[u^2+1]^2} = \ln\left| x\sqrt{x^2-x+1}\right| +$$

$$\frac{7\sqrt{3}}{3}\operatorname{arctg} u - \frac{1}{2(x^2-x+1)} - \frac{44\sqrt{3}}{9}\left(\frac{1}{2}\operatorname{arctg} u + \frac{1}{2}\frac{u}{u^2+1}\right) + C =$$

$$\ln\left| x\sqrt{x^2-x+1}\right| + \frac{7\sqrt{3}}{3}\operatorname{arctg}\frac{2x-1}{\sqrt{3}} -$$

$$\frac{22\sqrt{3}}{9}\operatorname{arctg}\frac{2x-1}{\sqrt{3}} - \frac{1}{2(x^2-x+1)} - \frac{22\sqrt{3}}{9}\frac{\frac{2x-1}{\sqrt{3}}}{\left(\frac{2x-1}{\sqrt{3}}\right)^2+1} + C =$$

$$\ln\left| x\sqrt{x^2-x+1}\right| - \frac{\sqrt{3}}{9}\operatorname{arctg}\frac{2x-1}{\sqrt{3}} - \frac{1}{3}\frac{11x-4}{x^2-x+1} + C.$$

□

6.B.30. Integrate

(a) $\int \frac{x}{1+x^4} \, dx$, $x \in \mathbb{R}$;

(b) $\int \frac{5\ln x}{x\ln^3 x + x\ln^2 x - 2x} \, dx$, $x > 0, x \neq$ e.

Solution. Case(a). The advantage of the method of integrating rational functions described above is its universality (using it, we can find primitive functions of every rational function). Sometimes though, using the substitution method or integrating by parts is more convenient. For example,

$$\int \frac{x}{1+x^4} \, dx = \left| \begin{array}{l} y = x^2 \\ dy = 2x\,dx \end{array} \right| = \int \frac{dy}{2(1+y^2)} = \frac{1}{2}\int \frac{dy}{1+y^2} =$$
$$\frac{1}{2}\operatorname{arctg} y + C = \frac{1}{2}\operatorname{arctg} x^2 + C.$$

Case (b). Using substitution, we obtain an integral of a rational function

$$\int \frac{5\ln x}{x\ln^3 x + x\ln^2 x - 2x} = \int \frac{5\ln x}{\ln^3 x + \ln^2 x - 2} \cdot \frac{1}{x} \, dx =$$

$$\left| \begin{array}{l} y = \ln x \\ dy = \frac{1}{x}\,dx \end{array} \right| = \int \frac{5y}{y^3+y^2-2} \, dy = \int \frac{1}{y-1} + \frac{-y+2}{y^2+2y+2} \, dy =$$

$$\int \frac{1}{y-1} \, dy - \frac{1}{2}\int \frac{2y+2}{y^2+2y+2} \, dy + 3\int \frac{1}{(y+1)^2+1^2} \, dy = \ln|y -$$

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.

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.[5]

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)

$$z^{(n)}(0) = P_n\big(g(z(0)), \dots, g^{(n-1)}(z(0))\big)$$
$$\geq P_n\big(|f(y(0))|, \dots, |f^{(n-1)}(y(0))|\big) \geq |y^{(n)}(0)|$$

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

---

[5]Although we shall not need the explicit formulae for these polynomials, they are well known under the name *Faá di Bruno's formula*. In principle, they are direct generalization of the Leibniz rule to higher order derivatives.

$1| - \frac{1}{2} \ln \left(y^2 + 2y + 2\right) + 3 \operatorname{arctg} (y + 1) + C = \ln |\ln x -$

$1| - \frac{1}{2} \ln \left(\ln^2 x + 2 \ln x + 2\right) + 3 \operatorname{arctg} (\ln x + 1) + C.$

$\square$

For an arbitrary function f that is continuous and bounded on a bounded interval (a, b), the so called Newton-Leibniz formula

$$(1) \qquad \int_a^b f(x)\, dx = [F(x)]_a^b := \lim_{x \to b-} F(x) - \lim_{x \to a+} F(x)$$

holds, where $F'(x) = f(x)$, $x \in (a, b)$. Emphasise that under the given conditions, the primitive function F always exists and so do both proper limits in (1). Hence to compute the definite integral, we only need to find the antiderivative and determine the respective one-sided limits (eventually only values of the function, if the primitive function is continuous at the boundary points of the interval).

6.B.31. Determine

(a) $\int \frac{1}{\sqrt{x^3} + \sqrt[5]{x^7}}\, dx$, $x > 0$;

(b) $\int \frac{x+1}{\sqrt[3]{3x+1}}\, dx$, $x \neq -\frac{1}{3}$;

(c) $\int \frac{1}{x} \sqrt{\frac{x+1}{x-1}}\, dx$, $x \in \mathbb{R} \smallsetminus [-1, 1]$;

(d) $\int \frac{1}{(x+4)\sqrt{x^2+3x-4}}\, dx$, $x \in (-\infty, -4) \cup (1, +\infty)$;

(e) $\int \frac{1}{1+\sqrt{-x^2+x+2}}\, dx$, $x \in (-1, 2)$;

(f) $\int \frac{1}{(x-1)\sqrt{x^2+x+1}}\, dx$, $x \neq 1$.

Solution. In this problem, we'll illustrate the use of the substitution method while integrating expressions containing roots.

Case (a). If the integral is in the form of

$$\int f\left(\sqrt[p(1)]{x}, \sqrt[p(2)]{x}, \ldots, \sqrt[p(j)]{x}\right) dx$$

for certain numbers $p(1), p(2), \ldots, p(j) \in \mathbb{N}$ and a rational function f (of more variables), the substitution $t^n = x$ is suggested, where n is the (least) common multiple of numbers $p(1), \ldots, p(j)$. Using this substitution, we can always reduce the integrand (the integrated function) to a rational function, which we can always integrate. We'll get

$$\int \frac{dx}{\sqrt{x^3}+\sqrt[5]{x^7}} = \int \frac{dx}{x\left(\sqrt{x}+\sqrt[5]{x^2}\right)} = \begin{vmatrix} t^{10} = x, & \sqrt[10]{x} = t \\ 10t^9\, dt = dx \end{vmatrix} =$$

$$\int \frac{10t^9}{t^{10}(t^5+t^4)}\, dt = 10 \int \frac{dt}{t^6+t^5} =$$

$$10 \int \left(\frac{1}{t} - \frac{1}{t^2} + \frac{1}{t^3} - \frac{1}{t^4} + \frac{1}{t^5} - \frac{1}{t+1}\right) dt =$$

$$10 \left[\ln t + \frac{1}{t} - \frac{1}{2t^2} + \frac{1}{3t^3} - \frac{1}{4t^4} - \ln(1+t)\right] + C =$$

$$\ln \frac{x}{(1+\sqrt[10]{x})^{10}} + \frac{10}{\sqrt[10]{x}} - \frac{5}{\sqrt[5]{x}} + \frac{10}{3\sqrt[10]{x^3}} - \frac{5}{2\sqrt[5]{x^2}} + C.$$

Case (b). For integrals

$$\int f\left(x, \sqrt[p(1)]{ax+b}, \sqrt[p(2)]{ax+b}, \ldots, \sqrt[p(j)]{ax+b}\right) dx,$$

otherwise). We may recognize the derivatives of a geometric series and write

$$(2) \qquad g(z) = C \sum_{n=0}^{\infty} \frac{z^n}{r^n} = C \frac{r}{r-z}$$

with derivatives $g^{(n)} = C \frac{n!}{r^n}$.

Finally, we have to prove that the solution of the equation $z' = g(z)$ is analytic. We can easily integrate this equation with separated variables directly. Written with the help of differentials, $(r - z)\, dz = Cr\, dx$. Thus, the implicit equation reads $\frac{1}{2}(r - z)^2 = -Crx + D$, 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. $\square$

6.2.16. 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 ∞ at the point $b = 2$. The integrand is continuous at all other points. Thus, for $0 < \delta < 2$, consider the integrals (substituting $y = 2 - x$)

$$I_\delta = \int_0^{2-\delta} \frac{dx}{\sqrt[4]{2 - x}} = \int_\delta^2 y^{-1/4}\, dy$$

$$= \left[\frac{4}{3} y^{3/4}\right]_\delta^2 = \frac{4}{3}[2^{3/4} - \delta^{3/4}].$$

Notice that $dy = -dx$. When $x = 2 - \delta$, $y = \delta$. When $x = 0$, $y = 2$.

The limit when $\delta \to 0$ 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*

where again $p(1), \ldots, p(j) \in \mathbb{N}$, f is a rational expression and $a, b \in \mathbb{R}$, we choose the substitution $t^n = ax + b$ while preserving the meaning of n. In this way, we'll get

$$\int \frac{x+1}{\sqrt[3]{3x+1}} \, dx = \left| \begin{matrix} t^3 = 3x+1 \\ dx = t^2 \, dt \end{matrix} \right| = \int \frac{\frac{t^3-1}{3}+1}{t} t^2 \, dt =$$

$$\int \frac{t^3-1+3}{3} t \, dt = \frac{1}{3} \int t^4 + 2t \, dt = \frac{1}{3} \left(\frac{t^5}{5} + t^2 \right) + C =$$

$$\frac{t^2}{3} \left(\frac{t^3}{5} + 1 \right) + C = \frac{\sqrt[3]{(3x+1)^2}}{3} \left(\frac{3x+1}{5} + 1 \right) + C =$$

$$\sqrt[3]{(3x+1)^2} \, \frac{x+2}{5} + C.$$

Case(c). Another generalizations are the integrals of the type

$$\int f \left(x, \sqrt[p(1)]{\frac{ax+b}{cx+d}}, \sqrt[p(2)]{\frac{ax+b}{cx+d}}, \ldots, \sqrt[p(j)]{\frac{ax+b}{cx+d}} \right) \, dx,$$

with the only additional condition on the values $a, b, c, d \in \mathbb{R}$ being $ad - bc \neq 0$. Preserving the meaning of the aforementioned symbols, we now put $t^n = \frac{ax+b}{cx+d}$. Specifically,

$$\int \frac{1}{x} \sqrt{\frac{x+1}{x-1}} \, dx = \left| \begin{matrix} t^2 = \frac{x+1}{x-1} \\ x = \frac{t^2+1}{t^2-1} \\ dx = -\frac{4t}{(t^2-1)^2} \, dt \end{matrix} \right| =$$

$$\int \frac{t^2-1}{t^2+1} \frac{-4t^2}{(t^2-1)^2} \, dt = \int \frac{-4t^2}{(t^2+1)(t^2-1)} \, dt =$$

$$\int \left(\frac{1}{t+1} - \frac{1}{t-1} - \frac{2}{t^2+1} \right) \, dt =$$

$$\ln |t+1| - \ln |t-1| - 2 \operatorname{arctg} t + C =$$

$$\ln \left| \sqrt{\frac{x+1}{x-1}} + 1 \right| - \ln \left| \sqrt{\frac{x+1}{x-1}} - 1 \right| - 2 \operatorname{arctg} \sqrt{\frac{x+1}{x-1}} + C.$$

The simplifications

$$\ln \left| \sqrt{\frac{x+1}{x-1}} + 1 \right| - \ln \left| \sqrt{\frac{x+1}{x-1}} - 1 \right| = \ln \left| \frac{\sqrt{\frac{x+1}{x-1}}+1}{\sqrt{\frac{x+1}{x-1}}-1} \right| =$$

$$\ln \left| \frac{\sqrt{\left|\frac{x+1}{x-1}\right|}+1}{\sqrt{\left|\frac{x+1}{x-1}\right|}-1} \right| = \ln \left| \frac{\sqrt{|x+1|}+\sqrt{|x-1|}}{\sqrt{|x+1|}-\sqrt{|x-1|}} \right| =$$

$$\ln \frac{\left(\sqrt{|x+1|} + \sqrt{|x-1|} \right)^2}{|\,|x+1|-|x-1|\,|} =$$

$$2 \ln \left(\sqrt{|x+1|} + \sqrt{|x-1|} \right) - \ln 2$$

for $x \in (-\infty, -1) \cup (1, \infty)$ then allow to write

$$\int \frac{1}{x} \sqrt{\frac{x+1}{x-1}} \, dx =$$

$$2 \ln \left(\sqrt{|x+1|} + \sqrt{|x-1|} \right) - 2 \operatorname{arctg} \sqrt{\frac{x+1}{x-1}} + C.$$

Cases (d), (e), (f). Now we'll focus on the integrals

$$\int f \left(x, \sqrt{ax^2+bx+c} \right) \, dx,$$

where we expect $a \neq 0$ and $b^2 - 4ac \neq 0$ for otherwise arbitrary numbers $a, b, c \in \mathbb{R}$. Recall that f is a rational expression. We'll distinguish two cases, when the quadratic polynomial $ax^2 + bx + c$ has real roots and when it doesn't.

If $a > 0$ and the polynomial $ax^2 + bx + c$ has real roots x_1, x_2, we'll use the representation

$$\sqrt{ax^2+bx+c} = \sqrt{a} \sqrt{(x-x_1)^2 \frac{x-x_2}{x-x_1}} =$$

$$\sqrt{a} \, |x-x_1| \, \sqrt{\frac{x-x_2}{x-x_1}}$$

integrals of the first kind. The integrals of unbounded functions on finite intervals are *improper Riemann integrals of the second kind.*

More explicitly, for $a \in \mathbb{R}$

$$I = \int_a^\infty f(x) \, dx = \lim_{b \to \infty} \int_a^b f(x) \, dx,$$

if the integrals and limit on the right hand side exist. Similarly we can have a finite upper bound an 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

$$\int_{-\infty}^\infty f(x) \, dx = \int_{-\infty}^a f(x) \, dx + \int_a^\infty f(x) \, dx.$$

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 $b \to \infty$ diverge to infinity.

The integrated functions may have more discontinuities with infinite one-sided limits. 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.

6.2.17. New acquisitions to the ZOO. It might seem that indefinite integrals can be described in terms of known elementary functions. But this is false.

On the contrary, 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 6.3.11 below).

In the methods of signal processing, the function

$$\operatorname{sinc}(x) = \frac{\sin(x)}{x}$$

is important (cf. the discussion of Fourier transform in 7.2.6). Check yourself that it is a smooth function with limit values

$$f(0) = 1, \; f'(0) = 0, \; f''(0) = -\frac{2}{3}.$$

This even function has its absolute maximum at the point $x = 0$. It oscillates with a fast decreasing amplitude as x approaches infinity.

and let $t^2 = \frac{x-x_2}{x-x_1}$. If $a < 0$ and the polynomial $ax^2 + bx + c$ has real roots $x_1 < x_2$, we'll use the representation

$$\sqrt{ax^2 + bx + c} = \sqrt{-a}\,\sqrt{(x-x_1)^2\,\tfrac{x_2-x}{x-x_1}} =$$
$$\sqrt{-a}\,(x-x_1)\,\sqrt{\tfrac{x_2-x}{x-x_1}}$$

and let $t^2 = \frac{x_2-x}{x-x_1}$. If the polynomial ax^2+bx+c doesn't have real roots (necessarily for $a > 0$), we choose the substitution

$$\sqrt{ax^2 + bx + c} = \pm\sqrt{a}\cdot x \pm t$$

with any choice of the signs. Note that we of course choose the signs so that we get as easy expression to integrate as possible. In all these cases, these substitutions again lead to rational functions.

Hence

(d)

$$\int \frac{dx}{(x+4)\sqrt{x^2+3x-4}} = \int \frac{dx}{(x+4)\sqrt{(x-1)(x+4)}} =$$
$$\int \frac{dx}{(x+4)|\,x+4\,|\sqrt{\frac{x-1}{x+4}}} = \left| \begin{array}{c} t^2 = \frac{x-1}{x+4} \\ x = \frac{5}{1-t^2} - 4 \\ dx = \frac{10t}{(1-t^2)^2}\,dt \end{array} \right| =$$
$$\int \frac{\frac{10t}{(1-t^2)^2}}{\left(\frac{5}{1-t^2}\right)\left|\frac{5}{1-t^2}\right|\,t}\,dt = \int \frac{2}{5}\frac{|\,1-t^2\,|}{1-t^2}\,dt =$$
$$\tfrac{2}{5}\operatorname{sgn}(1-t^2)\int 1\,dt = \tfrac{2}{5}\operatorname{sgn}\left(\tfrac{5}{x+4}\right)t + C =$$
$$\tfrac{2}{5}\operatorname{sgn}(x)\sqrt{\tfrac{x-1}{x+4}} + C;$$

(e)

$$\int \frac{dx}{1+\sqrt{-x^2+x+2}} = \int \frac{dx}{1+\sqrt{-(x-2)(x+1)}} =$$
$$\int \frac{dx}{1+(x+1)\sqrt{\frac{2-x}{x+1}}} = \left| \begin{array}{c} t^2 = \frac{2-x}{x+1} \\ x = \frac{3}{t^2+1} - 1 \\ dx = \frac{-6t}{(t^2+1)^2}\,dt \end{array} \right| =$$
$$\int \frac{\frac{-6t}{(t^2+1)^2}}{1+\frac{3}{t^2+1}\,t}\,dt = \int \frac{-6t}{(t^2+1)^2}\,\frac{t^2+1}{t^2+3t+1}\,dt =$$
$$\int \frac{-6t}{(t^2+1)(t^2+3t+1)}\,dt =$$
$$\int \left(-\tfrac{4}{5}\frac{\sqrt{5}}{2t+3+\sqrt{5}} - \tfrac{2}{5}\frac{2}{t^2+1} - \tfrac{4}{5}\frac{\sqrt{5}}{-2t-3+\sqrt{5}} \right)dt =$$
$$-\tfrac{2\sqrt{5}}{5}\ln\left|\,2t+3+\sqrt{5}\,\right| - 2\operatorname{arctg} t +$$
$$\tfrac{2\sqrt{5}}{5}\ln\left|\,-2t-3+\sqrt{5}\,\right| + C =$$
$$-\tfrac{2\sqrt{5}}{5}\ln\left|\,2\sqrt{\tfrac{2-x}{x+1}}+3+\sqrt{5}\,\right| - 2\operatorname{arctg}\sqrt{\tfrac{2-x}{x+1}} +$$
$$\tfrac{2\sqrt{5}}{5}\ln\left|\,-2\sqrt{\tfrac{2-x}{x+1}}-3+\sqrt{5}\,\right| + C =$$
$$\tfrac{2\sqrt{5}}{5}\ln\frac{2\sqrt{\frac{2-x}{x+1}}+3-\sqrt{5}}{2\sqrt{\frac{2-x}{x+1}}+3+\sqrt{5}} - 2\operatorname{arctg}\sqrt{\tfrac{2-x}{x+1}} + C;$$

(f)

$$\int \frac{dx}{(x-1)\sqrt{x^2+x+1}} = \left| \begin{array}{c} \sqrt{x^2+x+1} = x+t \\ x^2+x+1 = x^2+2xt+t^2 \\ x = -\frac{t^2+2t-2}{2t-1}+1 \\ dx = \frac{-2(t^2-t+1)}{(2t-1)^2}\,dt \end{array} \right| =$$
$$\int \frac{\frac{-2(t^2-t+1)}{(2t-1)^2}}{-\frac{t^2+2t-2}{2t-1}\,\frac{t^2-t+1}{2t-1}}\,dt = \int \frac{2}{t^2+2t-2}\,dt =$$

The *sine integral function* is defined by

$$\operatorname{Si}(x) = \int_0^x \operatorname{sinc}(t)\,dt.$$

Other important functions are *Fresnel's sine and cosine integrals*

$$\operatorname{FresnelS}(x) = \int_0^x \sin\left(\tfrac{1}{2}\pi t^2\right)dt$$
$$\operatorname{FresnelC}(x) = \int_0^x \cos\left(\tfrac{1}{2}\pi t^2\right)dt.$$

The function $Si(x)$ is shown in the left figure. Both Fresnel's functions are shown on the right.
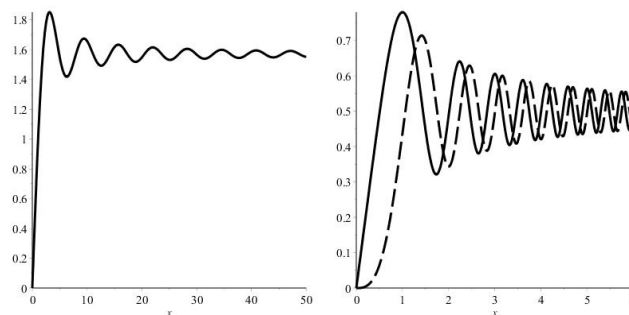
An even more important way how to get new functions is to add some free parameter in the integral. One of the most important mathematical functions ever is the *Gamma function*. It is defined for all positive real numbers z by

$$\Gamma(z) = \int_0^\infty e^{-t}\,t^{z-1}\,dt.$$

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 function $f(x) = \ln(\Gamma(x+1))$.

$$\int \left(\frac{\sqrt{3}}{3} \frac{1}{t+1-\sqrt{3}} - \frac{\sqrt{3}}{3} \frac{1}{t+1+\sqrt{3}} \right) dt =$$

$$\frac{\sqrt{3}}{3} \ln \left| t + 1 - \sqrt{3} \right| - \frac{\sqrt{3}}{3} \ln \left| t + 1 + \sqrt{3} \right| + C =$$

$$\frac{\sqrt{3}}{3} \ln \left| \frac{t+1-\sqrt{3}}{t+1+\sqrt{3}} \right| + C = \frac{\sqrt{3}}{3} \ln \left| \frac{\sqrt{x^2+x+1}-x+1-\sqrt{3}}{\sqrt{x^2+x+1}-x+1+\sqrt{3}} \right| + C.$$

$\square$

6.B.32. 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'll rather use a method we introduced for the complex roots case. That's because this method yields a very simple integral of a rational function, as can be seen from the calculation

$$\int \frac{dx}{x+\sqrt{x^2+x-1}} = \begin{vmatrix} \sqrt{x^2+x-1} = x+t \\ x^2+x-1 = x^2+2xt+t^2 \\ x = \frac{t^2+1}{1-2t} \\ dx = \frac{-2t^2+2t+2}{(1-2t)^2} \, dt \end{vmatrix} =$$

$$\int \frac{-2t^2+2t+2}{(t+2)(1-2t)} \, dt =$$

$$\int \left(1 - \frac{2}{t+2} - \frac{1}{2} \frac{1}{t-\frac{1}{2}} \right) dt =$$

$$t - 2\ln |t+2| - \frac{1}{2} \ln \left| t - \frac{1}{2} \right| + C =$$

$$\sqrt{x^2+x-1} - x - 2\ln \left(\sqrt{x^2+x-1} - x + 2 \right) -$$

$$\frac{1}{2} \ln \left| \sqrt{x^2+x-1} - x - \frac{1}{2} \right| + C.$$

Note that each recommended substitution (see the above problems) can be in most specific problems usually replaced by another substitution, which allows to obtain the result in a much easier way. 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.B.33. For $x > 0$ determine

(a) $\int \frac{(2+5x)^3}{\sqrt[4]{x^3}} \, dx$;

(b) $\int \frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}} \, dx$;

(c) $\int \frac{1}{\sqrt[4]{1+x^4}} \, dx$.

Solution. All three given integrals are binomial, i.e. they can be written as

$$\int x^m (a+bx^n)^p \, dx \quad \text{for some } a, b \in \mathbb{R}, \, m, n, p \in \mathbb{Q}.$$

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 a 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 $a+bx^n = t^s x^n$,
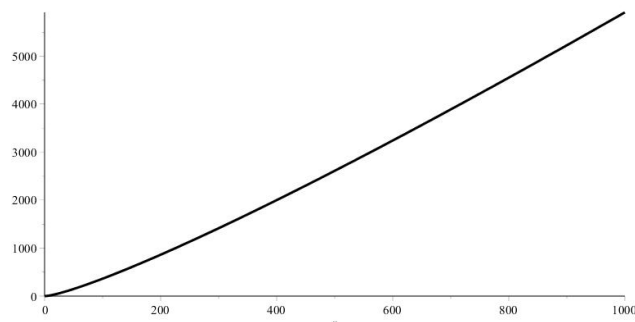
If we draw the function $x \ln x - x$ instead, there is not much difference to be seen. Hence it seems that the factorial $n!$ grows similarly to $e^{n \ln n - n} = n^n e^{-n}$. This is the famous *Stirling's approximation* formula. More precisely, one can verify

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \le n! \le e \, n^{n+\frac{1}{2}} e^{-n}.$$

In order to understand the qualitative behavior of such functions (e.g. their differentiability) 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 (invariance with respect to Euclidean motions and additivity with respect to finite union of disjoint objects). Open questions include: What are the "measurable objects" in the real line? How can the concept of area and volume be extended to higher dimensions?

The questions are answered partially with the help of the Riemann integral. First, measure the "volume" of one-dimensional subsets.

We say that the subset $A \subset \mathbb{R}$ is *(Riemann) measurable*, if the function $\chi : \mathbb{R} \to \mathbb{R}$

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

is Riemann integrable. That is, the (improper) integral

$$m(A) = \int_{-\infty}^{\infty} \chi_A(x) \, dx$$

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)

$$\int \frac{(2+5x)^3}{\sqrt[4]{x^3}}\,dx = \int x^{-\frac{3}{4}}(2+5x)^3\,dx =$$

$$\begin{vmatrix} p \in \mathbb{Z} \\ x = t^4 \\ dx = 4t^3\,dt \end{vmatrix} = 4\int \left(2+5t^4\right)^3\,dt =$$

$$4\int \left(8 + 60t^4 + 150t^8 + 125t^{12}\right)\,dt =$$

$$4\left(8t + 12t^5 + \frac{50}{3}t^9 + \frac{125}{13}t^{13}\right) + C =$$

$$4\left(8\sqrt[4]{x} + 12\sqrt[4]{x^5} + \frac{50}{3}\sqrt[4]{x^9} + \frac{125}{13}\sqrt[4]{x^{13}}\right) + C;$$

(b)

$$\int \frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}}\,dx = \int x^{-\frac{1}{2}}\left(1 + x^{\frac{1}{4}}\right)^{\frac{1}{3}}\,dx =$$

$$\begin{vmatrix} p \notin \mathbb{Z},\ \frac{m+1}{n} \in \mathbb{Z} \\ 1 + x^{\frac{1}{4}} = t^3 \\ x = (t^3 - 1)^4 \\ dx = 12t^2\left(t^3 - 1\right)^3\,dt \end{vmatrix} = 12\int t^3\left(t^3 - 1\right)\,dt =$$

$$12\int t^6 - t^3\,dt = 12\left(\frac{t^7}{7} - \frac{t^4}{4}\right) + C =$$

$$12\sqrt[3]{(1 + \sqrt[4]{x})^4}\left(\frac{1 + \sqrt[4]{x}}{7} - \frac{1}{4}\right) + C;$$

(c)

$$\int \frac{1}{\sqrt[4]{1+x^4}}\,dx = \int \left(1 + x^4\right)^{-\frac{1}{4}}\,dx =$$

$$\begin{vmatrix} p \notin \mathbb{Z},\ \frac{m+1}{n} \notin \mathbb{Z},\ \frac{m+1}{n} + p \in \mathbb{Z} \\ 1 + x^4 = t^4 x^4 \\ x = \left(t^4 - 1\right)^{-\frac{1}{4}} \\ dx = -t^3\left(t^4 - 1\right)^{-\frac{5}{4}}\,dt \end{vmatrix} = -\int \frac{t^2}{t^4 - 1}dt =$$

$$-\int \frac{t^2}{(t-1)(t+1)(t^2+1)}\,dt =$$

$$-\frac{1}{4}\int \left(\frac{1}{t-1} - \frac{1}{t+1} + \frac{2}{t^2+1}\right)\,dt =$$

$$-\frac{1}{4}\left(\ln|t-1| - \ln|t+1| + 2\operatorname{arctg}t\right) + C =$$

$$-\frac{1}{4}\left[\ln \frac{\sqrt[4]{\frac{1}{x^4}+1}-1}{\sqrt[4]{\frac{1}{x^4}+1}+1} + 2\operatorname{arctg}\left(\sqrt[4]{\frac{1}{x^4}+1}\right)\right] + C.$$

$\square$

6.B.34. For $x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, integrate

(a) $\int \frac{\sin^3 x}{1 + 4\cos^2 x + 3\sin^2 x}\,dx$;

(b) $\int \frac{1}{1 + \sin^2 x}\,dx$;

(c) $\int \frac{1}{2 - \cos x}\,dx$.

Solution. Integrals in the form of $\int f\left(\sin x, \cos x\right)\,dx$ for some rational function f are usually solved by the substitution method. If $f(\sin x, -\cos x) = -f(\sin x, \cos x)$, we choose $t = \sin x$; if $f(-\sin x, \cos x) = -f(\sin x, \cos x)$, we choose $t = \cos x$; and if $f(-\sin x, -\cos x) = f(\sin x, \cos x)$, then $t = \operatorname{tg} x$. If none of these equalities

exists (the finiteness of its value doesn't matter). The function χ_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 the expected 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.

The upper Riemann integral of the characteristic set χ_A corresponds to the infimum of the sums of lengths of finitely many disjoint 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 proceed in the same way in higher dimensions when defining the *Jordan measure*. For the definition of area (volume) in higher-dimensional space we generalize the concept of the Riemann integral,. We return to this in chapter 8. We remark here that the area of a plane figure bounded by a graph of a function in the way described above is consistent with expectations.

6.2.19. 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 (oriented according to the sign) over the interval $[a, b]$, which has the same area as that of the region between the x axis and the graph of the $f(x)$. 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*

$$\int_a^b f(x)\,dx = m(f)(b - a).$$

6.2.20. Length of a space 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 $F : \mathbb{R} \to \mathbb{R}^2$,

hold, the substitution $t = \text{tg}\, \frac{x}{2}$ is used. We'll show it on the given integrals.

Case (a). In the denominator, we have

$$1 + 4\cos^2 x + 3\sin^2 x = 4 + \cos^2 x$$

and in the numerator only the sine function to an odd power, i.e. the substitution $t = \cos x$, where $dt = -\sin x\, dx$, allows to replace all the sines and cosines and thus obtain

$$\int \frac{\sin^3 x}{1+4\cos^2 x + 3\sin^2 x}\, dx = \int \frac{\sin x (1 - \cos^2 x)}{4+\cos^2 x}\, dx =$$
$$\int \frac{-(1-t^2)}{4+t^2}\, dt = \int \left(1 - \frac{5}{4+t^2}\right) dt = t - \frac{5}{2}\, \text{arctg}\, \frac{t}{2} + C =$$
$$\cos x - \frac{5}{2}\, \text{arctg}\, \frac{\cos x}{2} + C.$$

Case (b). Because both the sine and cosine appear here to an even power, the substitution $t = \text{tg}\, x$ leads to

$$\sin^2 x = \frac{t^2}{1+t^2}, \quad \cos^2 x = \frac{1}{1+t^2}, \quad dx = \frac{1}{1+t^2}\, dt,$$

by which we obtain

$$\int \frac{dx}{1+\sin^2 x} = \int \frac{\frac{1}{1+t^2}}{1 + \frac{t^2}{1+t^2}}\, dt = \int \frac{1}{1+2t^2}\, dt =$$
$$\frac{\sqrt{2}}{2}\, \text{arctg}\, \left(\sqrt{2}t\right) + C = \frac{\sqrt{2}}{2}\, \text{arctg}\, \left(\sqrt{2}\, \text{tg}\, x\right) + C.$$

Case (c). Now we'll use the universal substitution $t = \text{tg}\, \frac{x}{2}$, where

$$\sin x = \frac{2t}{1+t^2}, \quad \cos x = \frac{1-t^2}{1+t^2}, \quad dx = \frac{2}{1+t^2}\, dt.$$

Then we can determine

$$\int \frac{dx}{2-\cos x} = \int \frac{\frac{2}{1+t^2}}{2 - \frac{1-t^2}{1+t^2}}\, dt = 2 \int \frac{dt}{1+3t^2} =$$
$$\frac{2\sqrt{3}}{3}\, \text{arctg}\, \left(\sqrt{3}t\right) + C = \frac{2\sqrt{3}}{3}\, \text{arctg}\, \left(\sqrt{3}\, \text{tg}\, \frac{x}{2}\right) + C.$$

$\square$

Definite integrals.

6.B.35. Compute the definite integrals

$$\int\limits_{\frac{\pi}{6}}^{\frac{\pi}{3}} \text{tg}^2 x\, dx, \qquad \int\limits_{0}^{\frac{\pi}{4}} \frac{x}{\cos^2 x}\, dx.$$

Solution. For $x \neq \frac{\pi}{2} + k\pi$, where $k \in \mathbb{Z}$, we have

$$\int \text{tg}^2 x\, dx = \text{tg}\, x - x + C,$$

as we have compute earlier. This implies that

$$\int\limits_{\pi/6}^{\pi/3} \text{tg}^2 x\, dx = [\text{tg}\, x - x]_{\pi/6}^{\pi/3} = \sqrt{3} - \frac{\pi}{3} - \left(\frac{1}{\sqrt{3}} - \frac{\pi}{6}\right) =$$
$$\frac{2}{\sqrt{3}} - \frac{\pi}{6}.$$

Of course, definite integrals can be also computed directly. For example, the substitution $y = \text{tg}\, x$ yields

$$F(t) = [f(t), g(t)]$$

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 Δs of the length of the curve corresponding to the increment Δt 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,

$$ds = \sqrt{1 + (y'(x))^2}\, dx,$$

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 2π, because π is defined in this way.

$$s = 2 \int_{-1}^{1} \sqrt{1 + (y')^2}\, dx = 2 \int_{-1}^{1} \sqrt{1 + \frac{x^2}{1 - x^2}}\, dx$$

$$= 2 \int_{-1}^{1} \frac{1}{\sqrt{1 - x^2}}\, dx = 2[\arcsin x]_{-1}^{1} = 2\pi.$$

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:

$$s(r) = 2 \int_{-r}^{r} \sqrt{1 + \frac{(x/r)^2}{1 - (x/r)^2}}\, dx = 2 \int_{-1}^{1} \frac{r}{\sqrt{1 - t^2}}\, dt$$

$$= 2r[\arcsin x]_{-1}^{1} = 2\pi r.$$

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 π is exactly the ratio, appearing in this dependency.

$$\int\limits_{\pi/6}^{\pi/3} \operatorname{tg}^2 x \, dx = \int\limits_{\pi/6}^{\pi/3} \frac{\sin^2 x}{\cos^2 x} \, dx =$$

$$\left| \begin{array}{l} y = \operatorname{tg} x; \, dy = \frac{dx}{\cos^2 x} \\ \sin^2 x = \frac{\operatorname{tg}^2 x}{1 + \operatorname{tg}^2 x} = \frac{y^2}{1 + y^2} \end{array} \right| = \int\limits_{1/\sqrt{3}}^{\sqrt{3}} \frac{y^2}{1 + y^2} \, dy =$$

$$\int\limits_{1/\sqrt{3}}^{\sqrt{3}} 1 - \frac{1}{1 + y^2} \, dy = [y - \operatorname{arctg} y]_{1/\sqrt{3}}^{\sqrt{3}} = \frac{2}{\sqrt{3}} - \frac{\pi}{6}.$$

When doing the substitution, we only need to not forget to change the limits of the integral to values gained by substituting $\sqrt{3} = \operatorname{tg}(\pi/3)$, $1/\sqrt{3} = \operatorname{tg}(\pi/6)$.

We'll compute the second integral by integration by parts for the definite integral. (Note that we also found the primitive function to function $y = x \cos^{-2} x$ earlier.) We have

$$\int\limits_0^{\pi/4} \frac{x}{\cos^2 x} \, dx = \left| \begin{array}{ll} F(x) = x & F'(x) = 1 \\ G'(x) = \frac{1}{\cos^2 x} & G(x) = \operatorname{tg} x \end{array} \right| =$$

$$[x \operatorname{tg} x]_0^{\pi/4} - \int\limits_0^{\pi/4} \operatorname{tg} x \, dx = [x \operatorname{tg} x]_0^{\pi/4} + \int\limits_0^{\pi/4} \frac{-\sin x}{\cos x} \, dx =$$

$$[x \operatorname{tg} x]_0^{\pi/4} + [\ln(\cos x)]_0^{\pi/4} = \frac{\pi}{4} + \ln \frac{\sqrt{2}}{2} = \frac{\pi - 2 \ln 2}{4}.$$

$\square$

6.B.36. Compute the definite integrals

(a) $\int_0^1 \frac{x}{\sqrt{1 - x^2}} \, dx$;

(b) $\int_1^2 \frac{1}{\sqrt{x^2 - 1}} \, dx$;

(c) $\int_0^1 \left(\frac{e^x}{e^{2x} + 3} + \frac{1}{\cos^2 x} \right) \, dx$;

Solution. We have

(a)

$$\int\limits_0^1 \frac{x}{\sqrt{1 - x^2}} \, dx = \left| \begin{array}{l} y = 1 - x^2 \\ dy = -2x \, dx \end{array} \right| = -\int\limits_1^0 \frac{y^{-1/2}}{2} \, dy =$$

$$\int\limits_0^1 \frac{y^{-1/2}}{2} \, dy = \left[\sqrt{y} \right]_0^1 = 1;$$

(b)

$$\int\limits_1^2 \frac{dx}{\sqrt{x^2 - 1}} = \left| \begin{array}{l} z = x + \sqrt{x^2 - 1} \\ dz = \frac{\sqrt{x^2 - 1} + x}{\sqrt{x^2 - 1}} \, dx \end{array} \right| = \int\limits_1^{2 + \sqrt{3}} \frac{1}{z} \, dz =$$

$$[\ln z]_1^{2 + \sqrt{3}} = \ln \left(2 + \sqrt{3} \right);$$

(c)

$$\int\limits_0^1 \left(\frac{e^x}{e^{2x} + 3} + \frac{1}{\cos^2 x} \right) dx = \int\limits_0^1 \frac{e^x}{e^{2x} + 3} \, dx + \int\limits_0^1 \frac{1}{\cos^2 x} \, dx =$$

$$\left| \begin{array}{l} p = e^x \\ dp = e^x \, dx \end{array} \right| = \int\limits_1^e \frac{1}{p^2 + 3} \, dp + [\operatorname{tg} x]_0^1 =$$

$$\frac{1}{3} \int\limits_1^e \frac{1}{\left(\frac{p}{\sqrt{3}} \right)^2 + 1} \, dp + \operatorname{tg} 1 = \left| \begin{array}{l} s = \frac{p}{\sqrt{3}} \\ ds = \frac{1}{\sqrt{3}} \, dp \end{array} \right| =$$

$$\frac{\sqrt{3}}{3} \int\limits_{1/\sqrt{3}}^{e/\sqrt{3}} \frac{1}{s^2 + 1} \, ds + \operatorname{tg} 1 = \frac{\sqrt{3}}{3} [\operatorname{arctg} s]_{1/\sqrt{3}}^{e/\sqrt{3}} + \operatorname{tg} 1 =$$

$$\frac{\sqrt{3}}{3} \left(\operatorname{arctg} \frac{e\sqrt{3}}{3} - \frac{\pi}{6} \right) + \operatorname{tg} 1;$$

6.2.21. 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 $0 \le x \le r$ determines one quarter the area. Use the substitution $x = r \sin t$, $dx = r \cos t \, dt$ (using the corollary for I_2 in the paragraph 6.2.6) to obtain by symmetry

$$a(r) = 4 \int_0^r \sqrt{r^2 - x^2} \, dx = 4r^2 \int_0^{\pi/2} \cos^2 t \, dt$$

$$= 4r^2 \int_0^{\pi/2} \sin^2 t \, dt = \frac{1}{2} 4r^2 \int_0^{\pi/2} \cos^2 t + \sin^2 t \, dt$$

$$= 2r^2 \int_0^{\pi/2} dt = \pi r^2.$$

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 π.

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 2π 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 Δx causes the area to increase by the multiple of the length Δs 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

$$A = 2\pi \int_a^b y(t) \sqrt{(y'(t))^2 + (x'(t))^2} \, dt.$$

6.B.37. Prove that

$$\frac{\sqrt{2}}{20} \leq \int\limits_0^1 \frac{x^9}{\sqrt{1+x}}\ dx \leq \frac{1}{10}.$$

Solution. Because

$$0 \leq \frac{x^9}{\sqrt{2}} \leq \frac{x^9}{\sqrt{1+x}} \leq x^9, \quad x \in [0,1],$$

the geometric meaning of the definite integral implies

$$\frac{\sqrt{2}}{20} = \int\limits_0^1 \frac{x^9}{\sqrt{2}}\ dx \leq \int\limits_0^1 \frac{x^9}{\sqrt{1+x}}\ dx \leq \int\limits_0^1 x^9\ dx = \frac{1}{10}.$$

$\square$

6.B.38. Without symbols of differentiation and integration, express

$$\left(\int\limits_x^0 t^5 \ln\left(t+1\right)\ dt\right)', \quad x \in (-1,1),$$

if the differentiation is done with respect to x.

Solution. Integration is often thought of as the inverse operation to differentiation. In this problem, we'll use this "inverseness". The function

$$F(x) := \int\limits_0^x t^5 \ln\left(t+1\right)\ dt, \quad x \in (-1,1)$$

is clearly the antiderivative of function $f(x) := x^5 \ln\left(x+1\right)$ on interval $(-1,1)$, i.e. by differentiating it, we'll get exactly f. Hence

$$\left(\int\limits_x^0 t^5 \ln\left(t+1\right)\ dt\right)' = -\left(\int\limits_0^x t^5 \ln\left(t+1\right)\ dt\right)' = -x^5 \ln\left(x+1\right).$$

$\square$

Improper integrals.

6.B.39. Decide if

$$\int\limits_1^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\ dx \in \mathbb{R}.$$

Solution. The improper integral represents the area of the figure between the graph of a positive function

$$y = \frac{\operatorname{arctg} x}{x\sqrt{x}}, \quad x \geq 1$$

and the x axis (from the left, the figure is bounded by the line $x = 1$). Hence the integral is a positive real number, or equals $+\infty$. We know that

$$\frac{\pi}{4} \leq \operatorname{arctg} x \leq \frac{\pi}{2}, \quad x \in [1, +\infty).$$

But that implies

$$\frac{\pi}{2} = \frac{\pi}{4} \int\limits_1^{+\infty} x^{-\frac{3}{2}}\ dx \leq \int\limits_1^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\ dx \leq \frac{\pi}{2} \int\limits_1^{+\infty} x^{-\frac{3}{2}}\ dx = \pi,$$

When computing the volume of the same solid, then the increase of Δx 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

$$V(f) = \pi \int_a^b \left(f(x)\right)^2 dx.$$

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.

$$A_r = 2\pi \int_{-r}^r r\sqrt{1 - (x/r)^2}\ \frac{1}{\sqrt{1 - (x/r)^2}}\ dt$$
$$= 2\pi r \int_{-r}^r dt = 4\pi r^2$$
$$V_r = \pi \int_{-r}^r \left(r^2 - x^2\right) dx$$
$$= \pi \left[r^2 x - \frac{1}{3}x^3\right]_{-r}^r = \frac{4}{3}\pi r^3.$$

Similarly to the circle, the ball is also the object which has the smallest surface are among all with a the given volume. That is the reason why soap bubbles almost always assume this shape.

6.2.22. Integral criterion for series. Using the improper integral, we can also decide the question of convergence for a class of infinite series:

Theorem. *Let $\sum_{n=1}^\infty f(n)$ be a series such that the function $f : \mathbb{R} \to \mathbb{R}$ is positive and nonincreasing on the interval $\langle 1, \infty)$. Then this series converges if and only if the integral*

$$\int_1^\infty f(x)\,\mathrm{d}x.$$

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

$$s_k' = \sum_{n=2}^k f(n) \leq \int_1^k f(x)\,\mathrm{d}x \leq \sum_{n=1}^\infty f(n) = s_k,$$

because s_k' is the lower sum of the Riemann integral while s_k is the upper sum. Thus, the integral converges if and only the series does, as expected. $\square$

3. Sequences, series and limit processes

While building a menagerie of functions, we encountered power series, which extend the collection of all polynomials in a natural way, see 5.4.3. We obtain a class of analytic functions in this way. But we have yet to prove that power series are continuous functions. We show below that not only

i.e. in particular

$$\int\limits_{1}^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\ dx \in \mathbb{R}.$$

$\square$

The formula (1) can be also used in a case when the function f is unbounded or the interval (a, b) is unbounded. We speak of the so called improper integrals. For the improper integrals, the limits on the right hand side may be improper and may not exist at all. If one of the limits doesn't exist or we receive an expression $\infty - \infty$, it means that the integral doesn't exist ($\infty - \infty$ doesn't have a character of an indefinite expression in this case). We say the integral oscilates. In every other case, we have the result (recall that $\infty + \infty = +\infty$, $-\infty - \infty = -\infty$, $\pm\infty + a = \pm\infty$ for $a \in \mathbb{R}$).

6.B.40. Determine

(a) $\int\limits_{1}^{\infty} \sin x\ dx$;

(b) $\int\limits_{1}^{\infty} \frac{dx}{x^4+x^2}$;

(c) $\int\limits_{0}^{4} \frac{dx}{\sqrt{x}}$;

(d) $\int\limits_{-1}^{1} \frac{dx}{x^2}$.

Solution. Case (a). We can immediately determine

$$\int\limits_{1}^{\infty} \sin x\ dx = [-\cos x]_1^{\infty} = \lim_{x\to\infty} (-\cos x) + \cos 1.$$

Because the limit on the right hand side doesn't exist, the integral oscilates.

Cases (b), (c). Analogously, we can easily compute

$$\int\limits_{1}^{\infty} \frac{dx}{x^4+x^2} = \int\limits_{1}^{\infty} \frac{dx}{x^2(x^2+1)} = \int\limits_{1}^{\infty} \frac{1}{x^2} - \frac{1}{1+x^2}\ dx =$$
$$\left[-\frac{1}{x} - \operatorname{arctg} x\right]_1^{\infty} = \lim_{x\to\infty} \left(-\frac{1}{x} - \operatorname{arctg} x\right) + \frac{1}{1} + \operatorname{arctg} 1 =$$
$$0 - \frac{\pi}{2} + 1 + \frac{\pi}{4} = 1 - \frac{\pi}{4}$$

and even more easily

$$\int\limits_{0}^{4} \frac{dx}{\sqrt{x}} = [2\sqrt{x}]_0^4 = 4 - 0 = 4,$$

where the primitive function is continuous from the right side at the origin (thus the limit equals the value of the function).

Case (d). If we'd mindlessly compute

$$\int\limits_{-1}^{1} \frac{dx}{x^2} = \left[-\frac{1}{x}\right]_{-1}^{1} = -1 - 1 = -2,$$

we'd receive an obviously wrong result (a negative value while integrating a positive function). The reason why the Newton-Leibniz formula cannot be applied in this way is the

are they continuous but it is possible to differentiate and integrate a power series term by term.

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 nonnegative and smaller than one at all points $0 \le x \le \pi$, except for $x = \frac{\pi}{2}$, where the value is 1. Hence on the whole interval $[0, \pi]$, these functions converge to the function

$$f(x) = \lim_{n\to\infty} f_n(x) = \begin{cases} 0 & \text{for all } x \neq \frac{\pi}{2} \\ 1 & \text{for } x = \frac{\pi}{2}. \end{cases}$$

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.

discontinuity of the given function at the origin. But if we use the additivity rule

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

which always holds, if the integrals on the right hand side are sensible, we'll find the correct result

$$\int_{-1}^1 \frac{dx}{x^2} = \int_{-1}^0 \frac{dx}{x^2} + \int_0^1 \frac{dx}{x^2} = \left[-\tfrac{1}{x}\right]_{-1}^0 + \left[-\tfrac{1}{x}\right]_0^1 =$$

$$\lim_{x\to 0-}\left(-\tfrac{1}{x}\right) - 1 - 1 - \lim_{x\to 0+}\left(-\tfrac{1}{x}\right) = \infty - 2 + \infty = +\infty.$$

Note that the even character of function $y = x^{-2}$ also implies

$$\int_{-1}^1 \frac{dx}{x^2} = 2\int_0^1 \frac{dx}{x^2} = 2\cdot\infty = +\infty.$$

$\square$

6.B.41. Compute the definite integrals

(a) $\int_0^\infty \frac{1}{(x+2)^5}\,dx$;

(b) $\int_{-2}^2 \ln|x|\,dx$;

(c) $\int_1^\infty \frac{e^{-\sqrt{x}}}{\sqrt{x}}\,dx$;

(d) $\int_{-1}^0 \frac{e^{1/x}}{x^3}\,dx$;

(e) $\int_1^2 \frac{1}{x\ln x}\,dx$.

Solution. We have

(a)

$$\int_0^\infty \frac{dx}{(x+2)^5} = -\tfrac{1}{4}\left[(x+2)^{-4}\right]_0^\infty =$$

$$-\tfrac{1}{4}\left(\lim_{x\to\infty}(x+2)^{-4} - 2^{-4}\right) = -\tfrac{1}{4}\left(0 - \tfrac{1}{16}\right) = \tfrac{1}{64};$$

(b)

$$\int_{-2}^2 \ln|x|\,dx = \int_{-2}^0 \ln|x|\,dx + \int_0^2 \ln|x|\,dx =$$

$$2\int_0^2 \ln x\,dx = \left|\begin{array}{l} F(x) = \ln x \\ G'(x) = 1 \end{array}\right|\left.\begin{array}{l} F'(x) = \tfrac{1}{x} \\ G(x) = x \end{array}\right| =$$

$$2\left([x\ln x]_0^2 - \int_0^2 1\,dx\right) = 2\left([x\ln x]_0^2 - [x]_0^2\right) =$$

$$2\left(2\ln 2 - \lim_{x\to 0+}(x\ln x) - 2 + 0\right) = 4\ln 2 - 4;$$

(c)

$$\int_1^\infty \frac{e^{-\sqrt{x}}}{\sqrt{x}}\,dx = \left|\begin{array}{l} t = \sqrt{x} \\ dt = \frac{1}{2\sqrt{x}}\,dx \end{array}\right| = 2\int_1^\infty e^{-t}\,dt =$$

$$2\left[-e^{-t}\right]_1^\infty = -2\left(\lim_{t\to\infty}e^{-t} - e^{-1}\right) = \tfrac{2}{e};$$

(d)

$$\int_{-1}^0 \frac{e^{1/x}}{x^3}\,dx = \left|\begin{array}{l} u = 1/x \\ du = -\frac{1}{x^2}\,dx \end{array}\right| = -\int_{-1}^{-\infty} u\,e^u\,du =$$

$$\int_{-\infty}^{-1} u\,e^u\,du = \left|\begin{array}{l} F(u) = u \\ G'(u) = e^u \end{array}\right|\left.\begin{array}{l} F'(u) = 1 \\ G(u) = e^u \end{array}\right| =$$

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 $m = n^3$, $n = 1,\ldots,10$.
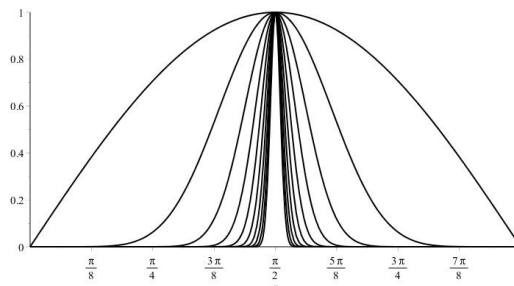
(2) We look at the second question, i.e. badly behaving derivatives. A natural idea on the same principle as above is to construct a sequence of functions which has the same nonzero derivative at one point, but becomes smaller and smaller. So they converge pointwise to the identically zero function.

The next figure plots the functions

$$f_n(x) = x(1-x^2)^n$$

on the interval $[-1,1]$ for values $n = m^2$, $m = 1,\ldots,10$.
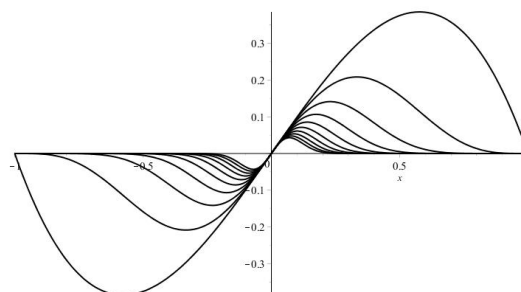
It is immediate that $\lim_{n\to\infty} f_n(x) = 0$ and that all functions $f_n(x)$ are smooth. Their derivative at $x = 0$ is

$$f'_n(0) = \left((1-x^2)^n - 2nx^2(1-x^2)^{n-1}\right)|_{x=0} = 1$$

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 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

$$[u\,e^u]_{-\infty}^{-1} - \int\limits_{-\infty}^{-1} e^u\,du = [u\,e^u]_{-\infty}^{-1} - [e^u]_{-\infty}^{-1} =$$

$$-\tfrac{1}{e} - \lim_{u\to-\infty} u\,e^u - \tfrac{1}{e} + \lim_{u\to-\infty} e^u = -\tfrac{2}{e};$$

(e)

$$\int\limits_1^2 \frac{dx}{x\ln x} = \left| \begin{array}{l} r = \ln x \\ dr = \tfrac{1}{x}\,dx \end{array} \right| = \int\limits_0^{\ln 2} \frac{dr}{r} = [\ln r]_0^{\ln 2} =$$

$$\ln(\ln 2) - \lim_{r\to 0+} \ln r = \ln(\ln 2) + \infty = +\infty.$$

□

6.B.42. Compute the improper integrals

$$\int\limits_0^\infty x^2\,e^{-x}\,dx; \qquad \int\limits_{-\infty}^\infty \frac{dx}{e^x + e^{-x}}.$$

Solution. Because the improper integral is a special case of a definite integral, we have at our disposal the basic methods to compute them. By integration by parts, we obtain

$$\int\limits_0^\infty x^2\,e^{-x}\,dx = \left| \begin{array}{ll} F(x) = x^2 & F'(x) = 2x \\ G'(x) = e^{-x} & G(x) = -e^{-x} \end{array} \right| =$$

$$\left[-x^2\,e^{-x}\right]_0^\infty + 2\int\limits_0^\infty x\,e^{-x}\,dx =$$

$$\left| \begin{array}{ll} F(x) = x & F'(x) = 1 \\ G'(x) = e^{-x} & G(x) = -e^{-x} \end{array} \right| =$$

$$- \lim_{x\to\infty} \tfrac{x^2}{e^x} + 2\left[-x\,e^{-x}\right]_0^\infty + 2\int\limits_0^\infty e^{-x}\,dx =$$

$$0 - 2\lim_{x\to\infty} \tfrac{x}{e^x} + 2\left[-e^{-x}\right]_0^\infty = 0 + 2\left(\lim_{x\to\infty} -e^{-x} + 1\right) = 2.$$

The substitution method then yields

$$\int\limits_{-\infty}^\infty \frac{dx}{e^x + e^{-x}} = \int\limits_{-\infty}^\infty \frac{e^x}{e^{2x}+1}\,dx = \left| \begin{array}{l} y = e^x \\ dy = e^x\,dx \end{array} \right| = \int\limits_0^\infty \frac{dy}{y^2+1} =$$

$$[\operatorname{arctg} y]_0^\infty = \lim_{y\to\infty} \operatorname{arctg} y = \tfrac{\pi}{2},$$

when the new limits of the integral are derived from the limits

$$\lim_{x\to-\infty} e^x = 0, \qquad \lim_{x\to\infty} e^x = +\infty.$$

□

6.B.43. Compute

$$\int\limits_0^\infty x^{2n+1}\,e^{-x^2}\,dx, \qquad n\in\mathbb{N}.$$

Solution. We'll first solve this problem by the substitution method and then repeatedly apply integration by parts, yielding

$$\int\limits_0^\infty x^{2n+1}\,e^{-x^2}\,dx = \left| \begin{array}{l} y = x^2 \\ dy = 2x\,dx \end{array} \right| = \tfrac{1}{2}\int\limits_0^\infty y^n\,e^{-y}\,dy =$$

$$\left| \begin{array}{ll} F(y) = y^n & F'(y) = ny^{n-1} \\ G'(y) = e^{-y} & G(y) = -e^{-y} \end{array} \right| =$$

$$\tfrac{1}{2}\left([-y^n\,e^{-y}]_0^\infty + n\int\limits_0^\infty y^{n-1}\,e^{-y}\,dy\right) =$$

$$\tfrac{n}{2}\int\limits_0^\infty y^{n-1}\,e^{-y}\,dy =$$

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 $n = m^2, m = 1,\ldots,10$ are on the next diagram.
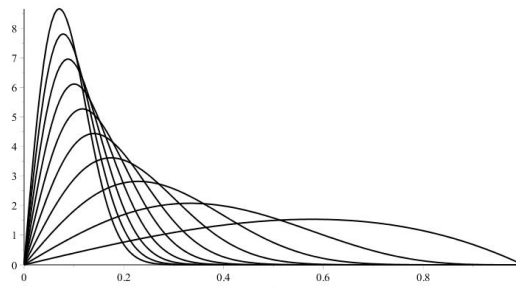
We verify that the values of these functions converge to zero for every $x \in [0,1]$ (for example $\ln(f_n(x)) \to -\infty$). But for all n

$$\int_0^1 f_n(x)\,dx = 1 \neq 0.$$

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 $f_n(x) \to f(x)$ 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 ε, 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 ε, it is independent on the point $x \in [a,b]$. This is the difference from pointwise convergence, where N depends on both ε 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 ε, 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 for uniform convergence (but beware the subtleties when differentiating).

$$\begin{vmatrix} F(y) = y^{n-1} & F'(y) = (n-1)y^{n-2} \\ G'(y) = \mathrm{e}^{-y} & G(y) = -\mathrm{e}^{-y} \end{vmatrix} =$$

$$\frac{n}{2}\left(\left[-y^{n-1}\,\mathrm{e}^{-y}\right]_0^\infty + (n-1)\int\limits_0^\infty y^{n-2}\,\mathrm{e}^{-y}\,dy \right) =$$

$$\frac{n(n-1)}{2}\int\limits_0^\infty y^{n-2}\,\mathrm{e}^{-y}\,dy = \cdots = \frac{n(n-1)\cdots 2}{2}\int\limits_0^\infty y\,\mathrm{e}^{-y}\,dy =$$

$$\begin{vmatrix} F(y) = y & F'(y) = 1 \\ G'(y) = \mathrm{e}^{-y} & G(y) = -\mathrm{e}^{-y} \end{vmatrix} =$$

$$\frac{n!}{2}\left(\left[-y\,\mathrm{e}^{-y}\right]_0^\infty + \int\limits_0^\infty \mathrm{e}^{-y}\,dy \right) = \frac{n!}{2}\left[-\mathrm{e}^{-y}\right]_0^\infty = \frac{n!}{2}.$$

$\square$

6.B.44. In dependancy on $a \in \mathbb{R}^+$ determine the integral $\int_0^1 \frac{1}{x^a}\,dx$. $\bigcirc$

Lengths, areas, surfaces, volumes.

6.B.45. Determine the length of the curve given parametrically:

$$x = \sin^2 t, \quad y = \cos^2 t,$$

for $t \in [0, \frac{\pi}{2}]$.

Solution. According to **??**, the length of a curve is given by the integral

$$\int_0^{\frac{\pi}{2}} \sqrt{(x'(t))^2 + (y'(t))^2}\,dt$$

$$= \int_0^{\frac{\pi}{2}} \sqrt{(\sin 2t)^2 + (-\sin 2t)^2}\,dt$$

$$= \int_0^{\frac{\pi}{2}} \sqrt{2}\sin 2t\,dt = \sqrt{2}.$$

If we realize that the given curve is a part of the line $y = 1 - x$ (since $\sin^2 t + \cos^2 t = 1$) and moreover the segment with boundary points $[0,1]$ (for $t = 0$) and $[1,0]$ (for $t = \frac{\pi}{2}$), we can immediately writw its length $\sqrt{2}$. $\square$

6.B.46. Determine the length of a curve given parametrically:

$$x = t^2, \quad y = t^3$$

for $t \in [0, \sqrt{5}]$.

Solution. We'll again determine the length l by using the formula **??**:

$$l = \int_0^{\sqrt{5}} \sqrt{4t^2 + 9t^4}\,dt = \int_0^{\sqrt{5}} t\sqrt{9t^2 + 4}\,dt$$

$$= \frac{1}{2}\int_0^5 \sqrt{9u + 4}\,dt = \frac{2}{27}[(9u+4)^{\frac{3}{2}}]_0^5 = \frac{335}{27}$$

$\square$

6.B.47. Determine the area to the right of the line $x = 3$ and bounded by the graph of a function $y = \frac{1}{x^3-1}$ and the x axis.

Solution. The are is given by the improper integral $\int_3^\infty \frac{1}{x^3-1} \, \mathrm{d}x$. We'll compute it using decomposition into partial fractions:

$$\frac{1}{x^3-1} = \frac{Ax+B}{x^2+x+1} + \frac{C}{x-1},$$
$$1 = (Ax+B)(x-1) + C(x^2+x+1),$$
$$x = 1 \implies C = \frac{1}{3},$$
$$x^0 : 1 = C - B \implies B = -\frac{2}{3},$$
$$x^2 : 0 = A + C \implies A = -\frac{1}{3}$$

and we can write

$$\int_3^\infty \frac{1}{x^3-1} \, \mathrm{d}x = \frac{1}{3} \int_3^\infty \left(\frac{1}{x-1} - \frac{x+2}{x^2+x+1} \right) \mathrm{d}x.$$

Now we'll seperately determine the indefinite integral $\int \frac{x+2}{x^2+x+1} \, \mathrm{d}x$:

$$\int \frac{x+2}{x^2+x+1} \, \mathrm{d}x$$
$$= \int \frac{x+\frac{1}{2}}{(x+\frac{1}{2})^2 + \frac{3}{4}} \, \mathrm{d}x + \frac{3}{2} \int \frac{1}{(x+\frac{1}{2})^2 + \frac{3}{4}} \, \mathrm{d}x$$

$$\left|\begin{array}{l} \text{substitution at the first integral} \\ t = x^2 + x + 1 \\ \mathrm{d}t = 2(x+\frac{1}{2}) \, \mathrm{d}x \end{array}\right|$$

$$= \frac{1}{2} \int \frac{1}{t} \, \mathrm{d}t + \frac{3}{2} \int \frac{1}{(x+\frac{1}{2})^2 + \frac{3}{4}}$$

$$\left|\begin{array}{l} \text{substitution at the first integral} \\ s = x + \frac{1}{2} \\ \mathrm{d}s = \mathrm{d}x \end{array}\right|$$

$$= \frac{1}{2} \ln(x^2+x+1) + \frac{3}{2} \int \frac{1}{s^2 + \frac{3}{4}} \, \mathrm{d}s$$

$$= \frac{1}{2} \ln((x^2+x+1) + \frac{3}{2} \frac{4}{3} \int \frac{1}{\left(\frac{2}{\sqrt{3}}s\right)^2 + 1} \, \mathrm{d}s$$

$$\left|\begin{array}{l} \text{substitution at the second integral} \\ u = \frac{2}{\sqrt{3}}s \\ \mathrm{d}u = \frac{2}{\sqrt{3}}s \, \mathrm{d}s \end{array}\right|$$

$$= \frac{1}{2} \ln(x^2+x+1) + 2\frac{\sqrt{3}}{2} \int \frac{1}{u^2+1} \, \mathrm{d}u$$

$$= \frac{1}{2} \ln(x^2+x+1) + \sqrt{3} \arctan(u)$$

$$= \frac{1}{2} \ln(x^2+x+1) + \sqrt{3} \arctan\left(\frac{2x+1}{\sqrt{3}}\right).$$

(3) Let $f_n(x)$ be a sequence of functions differentiable on a closed interval $[a, b]$ and assume $f_n(x_0) \to f(x_0)$ 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 $f(x) = \int_{x_0}^x g(t) \, dt$ is differentiable on the interval $[a, b]$, the functions $f_n(x)$ converge to $f(x)$ and $f'(x) = g(x)$. In other words,

$$\frac{d}{dx} f(x) = \frac{d}{dx} \left(\lim_{n \to \infty} f_n(x) \right) = \lim_{n \to \infty} \left(\frac{d}{dx} f_n(x) \right)$$

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 δ-neighbourhood of x_0. That is possible because $f_n(x)$ are continuous for all n. Then

$$|f(x) - f(x_0)| \le |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)|$$
$$+ |f_n(x_0) - f(x_0)| < 3\varepsilon$$

for all x in the δ-neighbourhood of x_0. This is the desired inequality with the bound 3ε. $\qquad\square$

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) $\lim_{x \to x_0} f_n(x) = a_n$, then the limit $\lim_{x \to x_0} f(x)$ exists if and only if the limit $\lim_{n \to \infty} a_n = a$ exists. Then they are equal, that is,

$$a = \lim_{n \to \infty} \left(\lim_{x \to x_0} f_n(x) \right) = \lim_{x \to x_0} \left(\lim_{n \to \infty} f_n(x) \right).$$

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.

In total, for the improper integral we can write:

$$\int_3^\infty \frac{1}{x^3 - 1}\, dx$$

$$= \frac{1}{3}\lim_{\delta\to\infty}\left[\ln|x-1|\right.$$

$$\left. -\frac{1}{2}\ln(x^2+x+1)-\sqrt{3}\arctan\left(\frac{2x+1}{\sqrt{3}}\right)\right]_3^\delta$$

$$= \frac{1}{3}\lim_{\delta\to\infty}\left(\frac{1}{3}\ln|\delta-1|\right.$$

$$\left. -\frac{1}{2}\ln(\delta^2+\delta+1)-\sqrt{3}\arctan\left(\frac{2\delta+1}{\sqrt{3}}\right)\right)$$

$$-\frac{1}{3}\ln 2+\frac{1}{6}\ln 13+\frac{\sqrt{3}}{3}\arctan\frac{7}{\sqrt{3}}$$

$$= \frac{1}{6}\ln 13+\frac{1}{\sqrt{3}}\arctan\frac{7}{\sqrt{3}}-\frac{1}{3}\ln 2-\frac{\sqrt{3}}{6}\pi.$$

$\square$

6.B.48. 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.

$$V = \int_0^2 \frac{x}{2}\, dx = \pi$$

$$S = 2\pi\int_0^2 \sqrt{\frac{x}{2}}\sqrt{1+\frac{1}{8x}}\, dx = 2\pi\int_0^2\sqrt{\frac{x}{2}+\frac{1}{16}}\, dx$$

$$= \pi\frac{17\sqrt{17}-1}{24}.$$

$\square$

6.B.49. 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

$$y = \tfrac{1}{\sqrt[3]{x-1}}.$$

Solution. First realize that

$$\tfrac{1}{\sqrt[3]{x-1}} < 0, \quad x \in [0,1), \qquad \tfrac{1}{\sqrt[3]{x-1}} > 0, \quad x \in (1,4]$$

and

$$\lim_{x\to 1-}\tfrac{1}{\sqrt[3]{x-1}} = -\infty, \quad \lim_{x\to 1+}\tfrac{1}{\sqrt[3]{x-1}} = +\infty.$$

The first part of the figure (below the x axis) is thus bounded by the curves

$$y = 0, \quad x = 0, \quad x = 1, \quad y = \tfrac{1}{\sqrt[3]{x-1}}$$

with an area given by the improper integral

Definition. The sequence of functions $f_n(x)$ on interval $[a, b]$ is *uniformly Cauchy*, if for every (small) positive number ε, there exists a large natural number N such that for all $x \in [a, b]$ and all $n \geq N$,

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Every uniformly convergent sequence of function 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 uniform 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 ε chosen beforehand and all $n \geq N$, $x \in [a, b]$. Now choose one such n and fix it, then

$$|f_n(x) - f(x)| = \lim_{m\to\infty}|f_n(x) - f_m(x)| \leq \varepsilon$$

for all $x \in [a, b]$. Hence the sequence $f_n(x)$ converges to its limit uniformly. $\square$

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 $f_n(x)$ 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. $\square$

$$S_1 = -\int_0^1 \frac{1}{\sqrt[3]{x-1}} \, dx;$$

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

$$\int \frac{1}{\sqrt[3]{x-1}} \, dx = \tfrac{3}{2} \sqrt[3]{(x-1)^2} + C,$$

the sum $S_1 + S_2$ can be gotten as

$$S = - \lim_{x \to 1-} \left(\tfrac{3}{2}\sqrt[3]{(x-1)^2} - \tfrac{3}{2} \right) +$$
$$\lim_{x \to 1+} \left(\tfrac{3}{2}\sqrt[3]{9} - \tfrac{3}{2}\sqrt[3]{(x-1)^2} \right) = \tfrac{3}{2}\left(1 + \sqrt[3]{9} \right).$$

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.B.50. Determine the avarage velocity v_p of a solid in the time interval $[1, 2]$, if its velocity is

$$v(t) = \frac{t}{\sqrt{1+t^2}}, \quad t \in [1, 2].$$

Omit the units.

Solution. To solve the problem, it suffices to realize that the sought avarage velocity is the mean value of function v on interval $[1, 2]$. Hence

$$v_p = \frac{1}{2-1} \int_1^2 \frac{t}{\sqrt{1+t^2}} \, dt = \int_2^5 \frac{1}{2\sqrt{x}} \, dx = \sqrt{5} - \sqrt{2},$$

with $1 + t^2 = x$, $t \, dt = dx/2$. $\square$

6.B.51. Compute the length s of a part of the curve calles tractrix given by the parametric description

$$f(t) = r\cos t + r\ln\left(\operatorname{tg} \tfrac{t}{2}\right), \quad g(t) = r\sin t, \quad t \in [\pi/2, a],$$

where $r > 0$, $a \in (\pi/2, \pi)$.

Solution. Since
$$f'(t) = -r\sin t + \frac{r}{2\operatorname{tg}\frac{t}{2}\cdot\cos^2\frac{t}{2}} = -r\sin t + \frac{r}{\sin t} = \frac{r\cos^2 t}{\sin t},$$

$$g'(t) = r\cos t \text{ on interval } [\pi/2, a],$$

for the length s we get

6.3.6. Proof of the third claim. For the corresponding result about derivatives, extra care is needed regarding the assumptions: If the 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 one of the assumptions 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, the corresponding results follow as a corollary in this way:

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:

$$\frac{d}{dx} \left(\sum_{n=1}^\infty f_n(x) \right) = \sum_{n=1}^\infty \frac{d}{dx} f_n(x).$$

$$s = \int_{\pi/2}^{a} \sqrt{\frac{r^2 \cos^4 t}{\sin^2 t} + r^2 \cos^2 t} \, dt = \int_{\pi/2}^{a} \sqrt{\frac{r^2 \cos^2 t}{\sin^2 t}} \, dt =$$

$$- r \int_{\pi/2}^{a} \frac{\cos t}{\sin t} \, dt = -r \left[\ln \left(\sin t\right)\right]_{\pi/2}^{a} = -r \ln \left(\sin a\right).$$

$\square$

6.B.52. 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.

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

$$\int_0^3 \pi f^2(x) \, dx = \pi \int_0^3 \sqrt{9x^2 - x^4} \, dx.$$

Using the substitution $t = \sqrt{9 - x^2}$ ($x \, dx = -t \, dt$), we can easily compute

$$\int_0^3 \sqrt{9x^2 - x^4} \, dx = \int_0^3 x \cdot \sqrt{9 - x^2} \, dx = -\int_3^0 t^2 \, dt = 9,$$

and receive the result 18π. $\square$

6.B.53. 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

$$V = \pi \int_a^{+\infty} \left(\frac{1}{x}\right)^2 \, dx = \pi \int_a^{+\infty} \frac{1}{x^2} \, dx =$$

$$\pi \left(\lim_{x \to +\infty} -\frac{1}{x} - \left(-\frac{1}{a}\right)\right) = \frac{\pi}{a}$$

and

$$S = 2\pi \int_a^{+\infty} \frac{1}{x} \cdot \sqrt{1 + \left(-\frac{1}{x^2}\right)^2} \, dx = 2\pi \int_a^{+\infty} \frac{\sqrt{x^4 + 1}}{x^3} \, dx$$

$$\geq 2\pi \int_a^{+\infty} \frac{1}{x} \, dx = 2\pi \left(\lim_{x \to +\infty} \ln x - \ln a\right) = +\infty.$$

The fact the the given solid (the so called Torricelli's trumper) cannot be painted with a finite amount of color, but can be filled with a finite amount of fluid, is called Torricelli's paradox. But realize that a real color painting has a

6.3.8. Test of uniform convergence. A simple way to test that a sequence of functions converges uniformly is to use a comparison with the absolute convergence of a suitable sequence of numbers. This is often called the *Weierstrass test*.

Suppose a sequence of functions $f_n(x)$ is given on an interval $I = [a, b]$ satifying

$$|f_n(x)| \leq a_n \in \mathbb{R}$$

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 $k > m$,

$$|s_k(x) - s_m(x)| = \left| \sum_{n=m+1}^{k} f_n(x) \right|$$

$$\leq \sum_{n=m+1}^{k} |f_n(x)| \leq \sum_{n=m+1}^{k} a_k.$$

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 has important results for power series

$$S(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$$

centered at a point x_0.

We saw earlier in 5.4.8, 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 ∞. (see 5.4.12).

In the proof of theorem 5.4.8, a comparison with a suitable geometric series is used to verify the convergence of the series $S(x)$. By the Weierstrass test, the 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 following crucial result is proved:

nonzero width, which the computation doesn't 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. □

C. Power series

6.C.1. Expand the function $\ln(1 + x)$ into a power series at point 0 and 1 and determine all $x \in \mathbb{R}$ for which these series converge.

Solution. First we'll determine the expansion at 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

$$[\ln(x + 1)]^{(n)} = (-1)^{n+1} \frac{(n - 1)!}{(x + 1)^n},$$

so after computing the derivatives at zero, we have $\ln(x + 1) = \ln 1 + \sum_{n=1}^{\infty} a_n x^n$, where

$$a_n = \frac{(-1)^{n+1}(n - 1)!}{n!} = \frac{(-1)^{n+1}}{n}.$$

Thus we can write

$$
\begin{aligned}
\ln(x + 1) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots \\
&= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n.
\end{aligned}
$$

For the radius of convergence, we can then use the limit of the quotient of the following coefficients of terms of the power series

$$r = \frac{1}{\lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right|} = \frac{1}{\lim_{n \to \infty} \frac{\frac{1}{n+1}}{\frac{1}{n}}} = 1.$$

Hence the series converges for arbitrary $x \in (-1, 1)$. For $x = -1$ we get the harmonic series (with a negative sign), 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, by computing the above derivatives from 6.C.1, we get

$$
\begin{aligned}
\ln(x + 1) = \ln(2) &+ \frac{1}{2}(x - 1) - \frac{1}{8}(x - 1)^2 \\
&+ \frac{1}{3 \cdot 2^3}(x - 1)^3 - \ldots \\
&= \ln(2) + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n \cdot 2^n}(x - 1)^n,
\end{aligned}
$$

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 nondifferentiable 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.

6.3.10. Laurent series. We return to the smooth function $f(x) = e^{-1/x^2}$ from paragraph 6.1.5 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:

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

[6]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.

and for the radius of convergence of this series, we get

$$r = \frac{1}{\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|} = \frac{1}{\lim_{n\to\infty} \frac{\frac{1}{2^{n+1}(n+1)}}{\frac{1}{2^n n}}} = 1.$$

The first series converges for $-1 < x \leq 1$, the second for $-1 < x \leq 3$. $\qquad \square$

6.C.2. Expand the function $\cos^2(x)$ into a power series (i.e. determine its Taylor expansion) at point 0 and determine for which real numbers this series converges. $\qquad \bigcirc$

6.C.3. Expand the function $\sin^2(x)$ into a power series at point 0 and determine for which real numbers this series converges $\qquad \bigcirc$

6.C.4. Expand the function $\ln(x^3 + 3x^2 + 3x + 1)$ into a power series at point 0 and determine for which $x \in \mathbb{R}$ it converges. $\qquad \bigcirc$

6.C.5. Expand the function $\ln \sqrt{x}$ into a power series at point 1 and determine for which $x \in \mathbb{R}$ it converges. $\qquad \bigcirc$

6.C.6. On the interval of convergence $(-1, 1)$, determine the sum of the series

$$\sum_{n=1}^{\infty} n(n+1)x^n.$$

Solution. We have

$$\sum_{n=1}^{\infty} n(n+1)x^n = \sum_{n=1}^{\infty} n\left(x^{n+1}\right)' = \left(\sum_{n=1}^{\infty} n\, x^{n+1}\right)' =$$

$$\left(\sum_{n=1}^{\infty} n\, x^{n-1}\, x^2\right)' = \left[x^2 \sum_{n=1}^{\infty} (x^n)'\right]' =$$

$$\left[x^2 \left(\sum_{n=1}^{\infty} x^n\right)'\right]' = \left[x^2 \left(-1 + \sum_{n=0}^{\infty} x^n\right)'\right]' =$$

$$\left[x^2 \left(-1 + \frac{1}{1-x}\right)'\right]' = \left[x^2 \cdot \frac{1}{(1-x)^2}\right]' = \frac{2x}{(1-x)^3}$$

for all $x \in (-1, 1)$. $\qquad \square$

6.C.7. For a 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}. $\qquad \bigcirc$

6.C.8. With an error lesser than $1/10$, approximately compute

$$\int_1^2 \left(x - \frac{\cos^{10} x}{10}\right) \ln x \, dx.$$

$\qquad \bigcirc$

6.C.9. Compute

$$\lim_{n\to\infty} \frac{\sqrt{1+\frac{1}{n}} + \sqrt{1+\frac{2}{n}} + \cdots + \sqrt{1+1}}{n}.$$

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

$$S(x) = \frac{a_{-s}}{(x-x_0)^r} + \cdots + \frac{a_{-1}}{x-x_0} + a_0 + a_1(x - x_0) + \ldots$$

$$= \sum_{n=-s}^{\infty} a_n(x - x_0)^n.$$

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 5.4.8 implies that its radius of convergence R is given by $R^{-1} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. 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 $r = \limsup_{n\to\infty} \sqrt[n]{|a_{-n}|}$.

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*

$$r = \limsup_{n\to\infty} \sqrt[n]{|a_{-n}|}, \quad R^{-1} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

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. Numerical approximation of integration. Just as in paragraph 6.1.16, 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 $a = x_0, x_1, \ldots, x_n = b$ 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 and the previous theorem. Integrate symmetrically around the middle values so

6.C.10. Aplications of the integral criterion of convergence. Now let's get back to (number) series. Thanks to the integral criterion of convergence (see 6.2.19), we can decide the question of convergence for a wider class of series: Decide, whether the following sums converge of diverge:

a) $\sum_{n=1}^{\infty} \frac{1}{n \ln n}$,

b) $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

Solution. First notice, that we cannot decide the convergence of none of these series by using the ratio or root test (all limits $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}|$ and $\lim_{n\to\infty} \sqrt[n]{a_n}$ equal 1). Using the integral criterion for convergence of series, we obtain:

a)
$$\int_1^\infty \frac{1}{x \ln(x)} \, dx = \int_0^\infty \frac{1}{t} \, dt = \lim_{\delta\to\infty} [\ln(t)]_0^\delta = \infty,$$
hence the given series diverges.

b)
$$\int_1^\infty \frac{1}{x^2} \, dx = \lim_{\delta\to\infty} \left[-\frac{1}{x} \right]_1^\delta = 1,$$
hence the given series converges.

$\square$

6.C.11. Using the integral criterion, decide the convergence of 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. $\square$

Uniform convergence.

6.C.12. Does the sequence of functions
$$y_n = e^{\frac{x^4}{4n^2}}, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}$$
converge uniformly on $\mathbb{R}$?

Solution. The sequence $\{y_n\}_{n\in\mathbb{N}}$ converges pointwise to the constant function $y = 1$ on $\mathbb{R}$, since

that the derivatives of odd orders cancel each other out while integrating:
$$\int_{-h/2}^{h/2} f(x_{i+1/2} + t) dt = \int_{-h/2}^{h/2} \left(\sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_{i+1/2}) t^n \right) dt$$
$$= \sum_{k=0}^{\infty} \left(\int_{-h/2}^{h/2} \frac{1}{k!} f^{(k)}(x_{i+1/2}) t^k \, dt \right)$$
$$= \sum_{k=0}^{\infty} \frac{h^{2k+1}}{2^{2k}(2k+1)!} f^{(2k)}(x_{i+1/2}).$$

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 $[x_i, 0]$, $[x_i, f_i]$, $[0, x_{i+1}]$, $[x_{i+1}, f_{i+1}]$ for approximation. This area is
$$P_i = \frac{1}{2}(f_i + f_{i+1})h$$
. In total, the integral I is approximated by
$$I_{\text{trap}} = \sum_{i=0}^{n-1} P_i = \frac{h}{2}(f_0 + 2f_1 + \cdots + 2f_{n-1} + f_n).$$

Compare I_{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:
$$f_{i+1/2\pm1/2} = f_{i+1/2} \pm \frac{h}{2} f'_{i+1/2} + \frac{h^2}{2!2^2} f''(i+1/2)$$
$$\pm \frac{h^3}{3!2^3} f^{(3)}(i+1/2) + \ldots.$$
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 integral 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
$$f_{i+1} = f(x_i + h) = f_i + \alpha_i h + \beta_i h^2$$
$$f_{i-1} = f(x_i - h) = f_i - \alpha_i h + \beta_i h^2$$

$$\lim_{n\to\infty} e^{\frac{x^4}{4n^2}} = e^0 = 1, \quad x \in \mathbb{R}.$$

But the computation

$$y_n\left(\sqrt{2n}\right) = e > 2 \quad \text{for all } n \in \mathbb{N}$$

implies that it's not a uniform convergence. (In the definition of uniform convergence, it suffices to consider $\varepsilon \in (0,1)$.) $\square$

6.C.13. 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

$$f_n(x) = \frac{\sqrt{x}\cdot n}{n^4+x^2}, \quad x > 0, \quad n \in \mathbb{N},$$

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 $x \in \left(0, n^2/\sqrt{3}\right)$ and $f_n'(x) < 0$ for $x \in \left(n^2/\sqrt{3}, +\infty\right)$ 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.14. 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

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)} x^{n+1} = \sum_{n=1}^{\infty}\left(\frac{(-1)^{n+1}}{n}\int x^n\,dx\right)$$

$$= \boxed{\int \sum_{n=1}^{\infty}\left(\frac{(-1)^{n+1}}{n}x^n\right)dx}$$

$$= \int \sum_{n=1}^{\infty}\left((-1)^{n+1}\int x^{n-1}\,dx\right)dx$$

$$= \int\left(\int \sum_{n=1}^{\infty}(-x)^{n-1}\,dx\right)dx = \int\left(\int 1 - x + x^2 - x^3 + \right.$$

$$\cdots\,dx\Big)dx = \int\left(\int \frac{1}{1+x}\,dx\right)dx = \boxed{\int \ln(1+x) + C_1\,dx}.$$

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

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 α_i)

$$P_i = \int_{-h}^{h} f_i + \alpha_i t + \beta_i t^2\,dt = 2hf_i + \frac{2}{3}\beta_i h^3$$

$$= 2hf_i + \frac{2h}{6}(f_{i+1} + f_{i-1} - 2f_i)$$

$$= \frac{h}{3}(f_{i+1} + f_{i-1} + 4f_i).$$

This procedure is called *Simpson's rule*[7]. The entire integral is now approximated by

$$I_{\text{Simp}} = \frac{1}{3}h\Big(f_0 + 4\sum_{m=0}^{n-1} f_{2m+1} + 2\sum_{m=1}^{n-1} f_{2m} + f_{2n}\Big).$$

As with the trapezoidal rule above, the total error is estimated by

$$|I - I_{\text{Simp}}| = \frac{1}{180}(b-a)h^4|f^{(4)}| + \mathrm{O}(h^5),$$

where $|f^{(4)}|$ represents the upper bound for $|f^{(4)}(x)|$ over the interval of integration.

6.3.12. Integrals dependent on parameters. When integrating a function $f(x, y_1, \ldots, y_n)$ of 1 real variable x depending on further real parameters $y_1, \ldots, y_n$ with respect to the single variable x, the result is a function $F(y_1, \ldots, y_n)$ 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 $\delta >$ 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

---

[7]This way of approximating the integral is attributed to the English mathematician and inventor Thomas Simpson (1710-1761).

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = \ln(1+x) + C_1, \qquad x \in (-1,1).$$

The choice $x = 0$ then yields $0 = \ln 1 + C_1$, i.e. $C_1 = 0$. Next,

$$\int \ln(1+x)\,dx = \big| \text{ per partes } \big|$$

$$= \left| \begin{array}{cc} u = \ln(1+x) & u' = \frac{1}{1+x} \\ v' = 1 & v = x \end{array} \right|$$

$$= x \ln(1+x) - \int \frac{x}{1+x}\,dx = x\ln(1+x) - \int 1 - \frac{1}{1+x}\,dx$$

$$= x\ln(1+x) - x + \ln(1+x) + C_2$$

$$= (x+1)\ln(x+1) - x + C_2.$$

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 $x \in (-1,1)$:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)} x^{n+1} = (x+1)\ln(x+1) - x.$$

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:

$$\lim_{x\to -1^+} (x+1)\ln(x+1) - x = \lim_{t\to 0^+} t \ln t + 1$$

$$= \lim_{t\to 0^+} \frac{\ln t}{\frac{1}{t}} + 1 = \lim_{t\to 0^+} \frac{\frac{1}{t}}{-\frac{1}{t^2}} + 1$$

$$= \lim_{t\to 0^+} -t + 1 = 1.$$

Of course, the convergence of the series at points ± 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 $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$. $\qquad \square$

6.C.15. 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

$$\sum_{n=1}^{\infty} \frac{1}{n 2^n}.$$

We'll use the fact that $\int\limits_2^{\infty} \frac{dx}{x^{n+1}} = \frac{1}{n 2^n}$.

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

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:

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 ξ_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 ξ_i, first use the standard properties of the absolute value and then exploit the choice of W:

$$\left| \sum_{i=0}^{k-1} f(\xi_i, \bar{y})(x_{i+1} - x_i) - \sum_{i=0}^{k-1} f(\xi_i, y)(x_{i+1} - x_i) \right|$$

$$\leq \sum_{i=0}^{k-1} \big| f(\xi_i, \bar{y}) - f(\xi_i, y) \big| (x_{i+1} - x_i)$$

$$< \varepsilon (b - a).$$

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. $\qquad \square$

6.3.13. Integrating twice. The fact that the integral $F(y) = \int_a^b f(x,y)dx$ of a continuous function $f : [a,b] \times [c,d] \to \mathbb{R}$ in the plane is again a continuous function $F : [c,d] \to \mathbb{R}$ allows us to repeat the integration and write

$$(1) \quad I = \int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d \Big(\int_a^b f(x,y)\,dx \Big) dy.$$

The next theorem is the simplest version of the claim known as *Fubini theorem*.

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

$$\sum_{n=1}^{\infty} \frac{1}{n2^n} = \sum_{n=1}^{\infty} \int_2^{\infty} \frac{dx}{x^{n+1}} = \int_2^{\infty} \left(\sum_{n=1}^{\infty} \frac{1}{x^{n+1}} \right) dx$$

$$= \int_2^{\infty} \frac{1}{x(x-1)} \, dx = \lim_{\delta \to \infty} \int_2^{\delta} \frac{1}{x-1} - \frac{1}{x} \, dx$$

$$= \lim_{\delta \to \infty} [(\ln(\delta - 1) - \ln(\delta) - \ln(1) + \ln 2]$$

$$= \lim_{\delta \to \infty} \left[\ln\left(\frac{\delta - 1}{\delta} \right) \right] + \ln(2)$$

$$= \ln\left(\lim_{\delta \to \infty} \frac{\delta - 1}{\delta} \right) + \ln 2 = \ln 2$$

$\square$

6.C.16. 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:

$$\lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \to \infty} \frac{(n+1)2^{-(n+1)}}{n2^n} = \lim_{n \to \infty} \frac{1}{2} \frac{n+1}{n} = \frac{1}{2}.$$

In total, according to (6.3.7) (3), we have

$$\int_{\ln 2}^{\ln 3} f(x) \, dx = \int_{\ln 2}^{\ln 3} \sum_{n=1}^{\infty} ne^{-nx} = \sum_{n=1}^{\infty} \int_{\ln 2}^{\ln 3} ne^{-nx} \, dx$$

$$= \sum_{n=1}^{\infty} [-e^{-nx}]_{\ln 2}^{\ln 3} = \sum_{n=1}^{\infty} \left(\frac{1}{2^n} - \frac{1}{3^n} \right) = 1 - \frac{1}{2} = \frac{1}{2}.$$

$\square$

6.C.17. Determine the following limit (give reasons for the procedure of computation):

$$\lim_{n \to \infty} \int_0^{\infty} \frac{\cos\left(\frac{x}{n} \right)}{\left(1 + \frac{x}{n} \right)^n} \, dx.$$

Theorem. *Consider a continuous function $f : [a, b] \times [c, d] \to \mathbb{R}$ in the plane $\mathbb{R}^2$. The multiple integration (1) is well defined and does not depend on the order of integration, i.e.,*

$$I = \int_c^d \left(\int_a^b f(x, y) \, dx \right) dy = \int_a^b \left(\int_c^d f(x, y) \, dy \right) dx..$$

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 Rieman integrals in (1) exist, thus we may fix a sequence Ξ_k of partitions of the interval $[a, b]$ into k subinterval $[x_{i-1}, x_i]$ of equal size $1/k$ and with representatives $\xi_{i,k}$, $i = 1, \ldots, k$, and similarly for the interval $[c, d]$ with the subintervals $[y_{i-1}, y_i]$ and representatives $\eta_{j,k}$, $j = 1, \ldots, k$. Then we may write

$$I = \int_c^d \lim_{k \to \infty} \left(\sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b - a) \right) dy$$

If $\frac{1}{k} < \delta$, then

$$\left| \int_a^b f(x, y) \, dy - \sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b - a) \right|$$

$$\leq \sum_{i=1}^k \left| \int_{x_{i-1}}^{x_i} f(x, y) \, dx - f(\xi_{i,k}, y) \frac{1}{k} (b - a) \right|$$

$$\leq \sum_{i=1}^k \int_{x_{i-1}}^{x_i} \left| f(x, y) - f(\xi_{i,k}, y) \right| dx$$

$$\leq \varepsilon(b - a).$$

Thus, the convergence of

$$\lim_{k \to \infty} \sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b - a) = F(y) = \int_a^b f(x, y) \, dx$$

is uniform on $[c, d]$. In particular, we may swap the integral and the limit to obtain

$$I = \lim_{k \to \infty} \int_c^d \left(\sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b - a) \right) dy$$

$$= \lim_{k \to \infty} \sum_{i=1}^k \left(\int_c^d f(\xi_{i,k}, y) dy \right) \frac{1}{k} (b - a)$$

$$= \lim_{k, \ell \to \infty} \sum_{i=1}^k \sum_{j=1}^\ell f(\xi_i, \eta_j) \frac{1}{k} \frac{1}{\ell} (b - a)(d - c).$$

Clearly, the same result will appear if we swap the order of the integration. $\square$

Solution. First we'll determine $\lim_{n\to\infty} \frac{\cos(\frac{x}{n})}{(1+\frac{x}{n})^n}$. The sequence of these functions converges pointwise and we have

$$\lim_{n\to\infty} \frac{\cos(\frac{x}{n})}{\left(1+\frac{x}{n}\right)^n} = \frac{1}{\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n} \overset{(??)}{=} \frac{1}{e^x}$$

It can be shown that the given sequence converges uniformly. Then according to (6.3.5),

$$\lim_{n\to\infty} \int_0^\infty \frac{\cos\left(\frac{x}{n}\right)}{\left(1+\frac{x}{n}\right)^n}\, \mathrm{d}x = \int_0^\infty \left[\lim_{n\to\infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1+\frac{x}{n}\right)^n}\right]\mathrm{d}x$$
$$= \int_0^\infty \frac{1}{e^x} = 1$$

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.18. 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.19. Find the analytic function whose Taylor series is

$$x - \tfrac{1}{3}x^3 + \tfrac{1}{5}x^5 - \tfrac{1}{7}x^7 + \cdots,$$

for $x \in [-1, 1]$. ○

6.C.20. From the knowledge of the sum of a geometric series, derive the Taylor series of function

$$y = \tfrac{1}{5+2x}$$

centered at the origin. Then determine its radius of convergence. ○

6.C.21. Expand the function

$$y = \tfrac{1}{3-2x}, \quad x \in \left(-\tfrac{3}{2}, \tfrac{3}{2}\right)$$

to a Taylor series centered at the origin. ○

6.C.22. 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.23. Express the function $y = \mathrm{e}^x$ 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.24. Find a function f such that for $x \in \mathbb{R}$, the sequence of functions

$$f_n(x) = \frac{n^2 x^3}{n^2 x^2 + 1}, \quad n \in \mathbb{N}$$

6.3.14. 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 $c \in [c,d]$, 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 c, then $\frac{d}{dy}F(c)$ exists as well and

$$\frac{d}{dy}F(c) = \int_a^b \frac{d}{dy}f(x,y)_{|y=c}\, dx.$$

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

$$H(y) = \int_{y_0}^y G(z)\, dz = \int_{y_0}^y \left(\int_a^b \frac{d}{dz}f(x,z)\, dx\right)dz$$
$$= \int_a^b \left(\int_{y_0}^y \frac{d}{dz}f(x,z)\, dz\right)dx$$
$$= \int_a^b (f(x,y) - f(x,y_0))\, dx$$
$$= F(y) - F(y_0).$$

Finally, differentiating with respect to y yields

$$G(y) = \frac{d}{dy}H(y) = \frac{dF}{dy}(y),$$

as desired. □

6.3.15. 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

to it. Is this convergence uniform on $\mathbb{R}$? ◯

6.C.25. Does the series

$$\sum_{n=1}^{\infty} \frac{n\,x}{n^4+x^2}, \quad \text{kde} \quad x \in \mathbb{R},$$

converge uniformly on the whole real axis? ◯

6.C.26. 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.27. Approximate

(a) cosine of ten degress 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.28. 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.29. Find the analytic function whose Taylor series is

$$x - \tfrac{1}{3}\,x^3 + \tfrac{1}{5}\,x^5 - \tfrac{1}{7}\,x^7 + \cdots,$$

for $x \in [-1,1]$. ◯

6.C.30. From the knowledge of the sum of a geometric series, derive the Taylor series of function

$$y = \tfrac{1}{5+2x}$$

centered at the origin. Then determine its radius of convergence. ◯

6.C.31. Use the derivatives of functions $y = \operatorname{tg} x$ and $y = \operatorname{cotg} x$ to find the indefinite integrals of functions

(a) $y = \operatorname{cotg}^2 x, \quad x \in (0,\pi)$;

(b) $y = \frac{1}{\sin^2 x \, \cos^2 x}, \quad x \in \left(0, \frac{\pi}{2}\right)$.

◯

6.C.32. By repeated use of integration by parts, for all $x \in \mathbb{R}$ determine

(a) $\int x^2 \sin x\ dx$;

(b) $\int x^2\, e^x\, dx$.

◯

6.C.33. For example by using integration by parts, determine

can be arranged at the infinitesimal level by exchanging the differential dx for $\varphi(x)dx$ for some suitable function φ.

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 nondecreasing function g on a finite interval $[a, b]$. For every partition Ξ with representative ξ_i and points of the partition

$$a = x_0, x_1, \ldots, x_n = b$$

the *Riemann–Stieltjes integral sum* of function $f(x)$ as

$$S_\Xi = \sum_{i=1}^{n} f(\xi_i)\big(g(x_i) - g(x_{i-1})\big).$$

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 Ξ with norm smaller than δ,

$$|S_\Xi - I| < \varepsilon.$$

For example, choose $g(x)$ on interval $[0, 1]$ as a piecewise constant function with finitely many discontinuities $c_1, \ldots, c_k$ and "jumps"

$$\alpha_i = \lim_{x \to c_{i+}} g(x) - \lim_{x \to c_{i-}} g(x),$$

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

$$\int x \ln^2 x \; dx$$

for $x > 0$. ◯

6.C.34. Using integration by parts, determine

$$\int \left(2 - x^2\right) e^x \, dx$$

on the whole real line. ◯

6.C.35. Integrate

(a) $\int (2x + 5)^{10} \, dx, \quad x \in \mathbb{R}$;

(b) $\int \frac{1}{x \ln^2 x} \, dx, \quad x > 0$;

(c) $\int e^{-x^3} x^2 \, dx, \quad x \in \mathbb{R}$;

(d) $\int 15 \frac{\arcsin^2 x}{\sqrt{1-x^2}} \, dx, \quad x \in (-1, 1)$;

(e) $\int \frac{\ln x}{x} \, dx, \quad x > 0$;

(f) $\int \frac{\operatorname{arctg} \sqrt{x}}{\sqrt{x}(1+x)} \, dx, \quad x > 0$;

(g) $\int \frac{e^x}{e^{2x}+3} \, dx, \quad x \in \mathbb{R}$;

(h) $\int \sin \sqrt{x} \, dx, \quad x > 0$

by using the substitution method. ◯

6.C.36. For $x \in (0, 1)$, by using suitable substitutions, reduce the integrals

$$\int x^2 \sqrt{\frac{x}{1-x}} \, dx; \quad \int \frac{dx}{(x-1)\sqrt{x^2+x+1}}$$

to integrals of rational functions. ◯

6.C.37. For $x \in (-\pi/2, \pi/2)$ compute

$$\int \frac{dx}{1+\sin^2 x}$$

using the substitution $t = \operatorname{tg} x$. ◯

6.C.38. How many distinct primitive functions to function $y = \cos(\ln x)$ does there exist on the interval $(0, 10)$? ◯

6.C.39. Give an example of a function f on th interval $I = [0, 1]$ that doesn't have a primitive function on I. ◯

6.C.40. Using the Newton integral, compute

(a) $\int\limits_0^\pi \sin x \, dx$;

(b) $\int\limits_0^1 \operatorname{arctg} x \, dx$;

(c) $\int\limits_{-\pi/4}^{3\pi/4} \frac{\cos x}{1+\sin x} \, dx$;

(d) $\int\limits_{1/e}^{e} |\ln x| \, dx$. ◯

6.C.41. Compute

$$\int\limits_1^2 \frac{x}{\sqrt{1+x^2}} \, dx.$$

function $g(x)$ on a finite interval $[a, b]$ define its *variation* by

$$\operatorname{var}_a^b g = \sup_\Xi \sum_{i=1}^n |g(x_i) - g(x_{i-1})|,$$

where the supremum is taken over all partitions Ξ 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.

As in the discussion of the Riemann integral, we derive the following theorem. 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.

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.

6.3.16. 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 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

6.C.42. For arbitrary real numbers $a < b$ determine

$$\int\limits_a^b \operatorname{sgn} x \; dx.$$

Recall that $\operatorname{sgn} x = 1$, for $x > 0$; $\operatorname{sgn} x = -1$, for $x < 0$; and $\operatorname{sgn} 0 = 0$.

6.C.43. Compute the definite integral

$$\int\limits_0^1 \frac{x^3}{1+x^4} \; dx.$$

6.C.44. For example by repeated integration by parts, compute

$$\int\limits_0^{\pi/2} e^{2x} \cos x \; dx.$$

6.C.45. Determine

$$\int\limits_{-1}^1 x^2 \, e^{-x} \; dx.$$

6.C.46. Compute the integral

$$\int\limits_{-1}^1 \frac{x}{\sqrt{5-4x}} \; dx$$

using the substitution method.

6.C.47. Compute

(a) $\int\limits_1^{e^8} \frac{dx}{x\sqrt{1+\ln x}}$;

(b) $\int\limits_0^{\ln 2} \frac{x}{e^x} \; dx.$

6.C.48. Which of the positive numbers

$$p := \int\limits_0^{\pi/2} \cos^7 x \; dx, \quad q := \int\limits_0^{\pi} \cos^2 x \; dx$$

is bigger?

6.C.49. Determine the signs of these three numbers (values of integrals)

$$a := \int\limits_{-2}^2 x^3 \, 2^x \; dx; \quad b := \int\limits_0^{\pi} \cos x \; dx; \quad c := \int\limits_0^{2\pi} \frac{\sin x}{x} \; dx.$$

6.C.50. Order the numbers

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 δ on a finite interval $[a, b]$ is called a *gauge*. A partition Ξ of interval $[a, b]$ with representatives ξ_i is δ–*gauged*, if

$$\xi_i - \delta(\xi_i) < x_{i-1} \le \xi_i \le x_i < \xi_i + \delta(\xi_i)$$

for all i.

The norm δ 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 δ, a δ–gauged partition with representatives exist. 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 δ on $[a, b]$, denote by M the set of all points $x \in [a, b]$ such that a δ–gauged partition with representatives can be found on $[a, x]$. M is nonempty and bounded, thus it has a supremum s. If $s \ne b$, 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.

Now we can state the following generalization of the Riemann integral.

Call it the *K-integral*[8].

---

[8]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.

$$A := \int_0^{\pi/2} \cos x \, \sin^2 x \, dx, \quad B := \int_0^{\pi/2} \sin^2 x \, dx, \quad C :=$$

$$\int_{-1}^{1} -x^5 \, 5^x \, dx,$$

$$D := \int_{2\pi}^{10} \frac{x^2+2}{x^6+4} \, dx + \int_{\pi}^{2\pi} \frac{x^2+2}{x^6+4} \, dx + \int_{10}^{\pi} \frac{x^2+2}{x^6+4} \, dx$$

by size. ◯

6.C.51. By considering the geometric meaning of the definite integral, determine

(a) $\int_{-2}^{2} |x-1| \, dx$;

(b) $\int_{-0,10}^{0,10} \operatorname{tg} x \, dx$;

(c) $\int_{0}^{2\pi} \sin x \, dx$.

◯

6.C.52. Compute $\int_{-1}^{1} |x| \, dx$. ◯

6.C.53. Determine

$$\int_{-1}^{1} x^5 \, \sin^2 x \, dx.$$

◯

6.C.54. Without using the symbols of differentiating and integrating, express

$$\left(\int_{x^2}^{a} 3t^2 \cos t \, dt \right)'$$

with variable $x \in \mathbb{R}$ and a real constant a, if we differentiate with respect to x. ◯

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 δ such that for every δ–gauged partition with representatives Ξ, the inequality $|S_\Xi - I| < \varepsilon$ is true for the corresponding Riemann sum S_Ξ.

6.3.17. 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 δ–gauged.

The Kurweil integral behaves as anticipated with respect to the sets of zero measure:

Theorem. *Let the function f, which is zero almost everywhere be defined on the interval $[a,b]$. Then the K-integral $\int_a^b f(x) d(x)$ 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 $M_k \subset [a,b]$, $k = 1, \ldots,$ for the subset of the points for which $k - 1 < |f(x)| \le k$. 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, δ is defined arbitrarily.

For any δ–gauged partition Ξ of the interval $[a,b]$ the bound on the corresponding Riemann sum is given as

$$\left| \sum_{j=0}^{n-1} f(\xi_n)(x_{i+1} - x_i) \right| = \left| \sum_{\substack{j=0 \\ \xi_i \in M}}^{n-1} f(\xi_n)(x_{i+1} - x_i) \right|$$

$$\le \sum_{k=1}^{\infty} \sum_{\substack{j=0 \\ \xi_i \in M_k}}^{n-1} \left| f(\xi_n) \right| (x_{i+1} - x_i)$$

$$\leq \sum_{k=1}^{\infty} k \left(\sum_{\substack{j=0 \\ \xi_i \in M_k}}^{n-1} m(J_{k,j}) \right).$$

To guarantee that this bound is smaller than a given ε, 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.*

6.3.18. 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.[9]

In 6.2.9, we deal with the relation between the derivatives $f(t)$ and the antiderviatives (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:

1ST AND 2ND FUNDAMENTAL THEOREMS OF CALCULUS

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 $F(t) - F(a)$.

6.3.19. 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

---

[9]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.

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

$$\lim_{t \to a-} \int_t^b f(x)\,dx$$

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 χ_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 $g(x) = |\operatorname{sinc}(x)|$ 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.

6.3.20. 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 $f_0(x) \le f_1(x) \le \dots$.

Theorem. *Suppose $f_0, f_1, f_2, \dots$ are all 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

$$\int_a^b f(t)\, dt = \lim_{n \to \infty} \int_a^b f_n(t)\, dt.$$

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. Extra examples for the whole chapter

6.D.1. Determine the significant properties of the function

$$f(x) = -\frac{x^2}{x+1}, \quad x \in \mathbb{R} \smallsetminus \{-1\}.$$

○

By "significant properties" is meant items such as domain, range, zeros, extrema, stationary and inflection points, points of discontinuity, intervals of increasing, decreasing, convexity, concavity, and asymptotes if applicable. Briefly, all the relevant information required to sketch a graph.

6.D.2. Determine the significant properties of the function

$$f(x) = \frac{1-x^3}{x^2}.$$

○

6.D.3. Determine the significant properties of the function

$$f(x) = \frac{x^3 - 3x^2 + 3x + 1}{x-1}.$$

○

6.D.4. Determine the significant properties of the function

$$f(x) = \sqrt[3]{x}\, e^{-x}.$$

○

6.D.5. Determine the significant properties of the function

$$f(x) = \operatorname{arctg} \frac{x}{2-x}.$$

○

6.D.6. Determine the significant properties of the function

$$\frac{\ln x}{x}.$$

Especially, find the extremes, the points of inflection and the asymptotes and sketch its graph.   ○

6.D.7. Determine the significant properties of the function

$$\ln(x^2 - 3x + 2) + x.$$

In particular, find the extremes, the points of inflection and the asymptotes: ○

6.D.8. Determine the significant properties of the function

$$(x^2 - 2)e^{x^2 - 1}.$$

In particular, find the extremes, the points of inflection and the asymptotes: ○

6.D.9. Determine the significant properties of the function

$$\ln(2x^2 - x - 1).$$

Among other things, find the extremes, the points of inflection and the asymptotes.   ○

6.D.10. Determine the significant properties of the function

$$\frac{x^2 - 2}{x - 1}.$$

Among other things, find the extremes, the points of inflection and the asymptotes.

6.D.11. Using any basic formulas, determine a primitive function for the function

(a) $y = \sqrt{x\sqrt{x\sqrt{x}}}, \quad x \in (0, +\infty)$;
(b) $y = (2^x + 3^x)^2, \quad x \in \mathbb{R}$;
(c) $y = \frac{1}{\sqrt{4-4x^2}}, \quad x \in (-1, 1)$;
(d) $y = \frac{\cos x}{1+\sin x}, \quad x \in \left(-\frac{\pi}{2}, \frac{3\pi}{2}\right)$.

6.D.12. Find a primitive function for the function
$$y = e^x + \frac{3}{\sqrt{4-x^2}}$$
on the interval $(-2, 2)$.

6.D.13. Determine
$$\int \frac{x^3}{1+x^4}\, dx, \quad x \in \mathbb{R}.$$

6.D.14. Determine
$$\int \frac{4}{x^2-2x+3}\, dx, \quad x \in \mathbb{R}.$$

6.D.15. For $x \in (0,1)$, compute
$$\int \left(\frac{x^2+1}{x(x^2-1)} + \frac{3}{\sqrt{4-4x^2}} + 4\sin x - 5\cos x\right) dx.$$

6.D.16. Determine the indefinite integrals

(a) $\int \operatorname{arctg} x\, dx, \quad x \in \mathbb{R}$;
(b) $\int \frac{\ln x}{x}\, dx, \quad x > 0$

using integration by parts.

6.D.17. Determine
$$\int \frac{\sqrt{x}}{\sqrt{x}+1}\, dx, \quad x > 0.$$

6.D.18. Compute

(a) $\int x^n \ln x\, dx, \quad x > 0, n \neq -1$;
(b) $\int \frac{x}{1+x^4}\, dx, \quad x \in \mathbb{R}$.

6.D.19. Determine

$$\int \frac{3x+5}{x^2+4x+8} \ dx, \quad x \in \mathbb{R}.$$

○

6.D.20. Compute the indefinite integral of the function
$$y = \frac{1}{(x^2+x+1)^2}, \quad x \in \mathbb{R}.$$

○

6.D.21. Determine
$$\int \frac{dx}{x^3+1}, \quad x \neq -1.$$

○

6.D.22. Integrate
$$\int \frac{1}{x^3-1} \ dx, \quad x \neq 1.$$

○

6.D.23. Compute the integral
$$\int \frac{x^3}{(x-1)(x-2)^2} \ dx, \quad x \in \mathbb{R} \smallsetminus \{1, 2\}.$$

○

6.D.24. For $x \in \left(0, \frac{\pi}{2}\right)$, compute

(a) $\int \sin^3 x \cos^4 x \ dx$;

(b) $\int \frac{1+\cos^2 x}{1+\cos 2x} \ dx$;

(c) $\int 2 \sin^2 \frac{x}{2} \ dx$;

(d) $\int \cos^2 x \ dx$;

(e) $\int \cos^5 x \sqrt{\sin x} \ dx$;

(f) $\int \frac{dx}{\sin^2 x \cos^4 x}$;

(g) $\int \frac{dx}{\sin^3 x}$;

(h) $\int \frac{dx}{\sin x}$.

○

6.D.25. Compute the indefinite integral

$$\int \frac{1}{x^4 + 3x^3 + 5x^2 + 4x + 2}\, dx.$$

6.D.26. Compute the integral

$$\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \frac{\sin t}{1 - \cos^2 t}\, dt.$$

6.D.27. Compute the integral

$$\int_0^{\ln 2} \frac{dx}{e^{2x} - 3e^x}.$$

6.D.28. Compute:

(i) $\int_0^{\frac{\pi}{2}} \sin x \sin 2x\, dx$,

(ii) $\int \sin^2 x \sin 2x\, dx$.

6.D.29. Compute the improper integral

(a) $\int\limits_{-\infty}^{+\infty} \frac{dx}{1+x^2}$;

(b) $\int\limits_{0}^{+\infty} \frac{dx}{x}$;

(c) $\int\limits_{0}^{4} \frac{2x^2+\sqrt{x}}{x}\, dx$;

(d) $\int\limits_{-1}^{1} \ln|x|\, dx$.

6.D.30. Determine

$$\int\limits_{0}^{3\pi/2} \frac{\cos x}{1+\sin x}\, dx.$$

6.D.31. Compute the improper integrals

$$\int\limits_{-\infty}^{+\infty} \frac{1}{x^2+x+1}\, dx.$$

6.D.32. Compute

$$\int\limits_{-\infty}^{+\infty} \frac{e^x}{e^{2x}+e^x+1}\, dx.$$

6.D.33. By using the substitution method, compute

$$\int\limits_{-\infty}^{0} x\,\mathrm{e}^{-x^2}\,dx; \quad \int\limits_{0}^{\infty} \frac{\mathrm{e}^{-(1/x)}}{x^2}\,dx.$$

6.D.34. Compute the integrals

$$\int\limits_{0}^{1} \frac{\mathrm{e}^{-\sqrt{x}}}{\sqrt{x}}\,dx; \quad \int\limits_{1}^{4} \frac{\mathrm{e}^{-\sqrt{x}}}{\sqrt{x}}\,dx; \quad \int\limits_{4}^{+\infty} \frac{\mathrm{e}^{-\sqrt{x}}}{\sqrt{x}}\,dx.$$

6.D.35. Find the values of $\alpha \in \mathbb{R}$, for which

(a) $\displaystyle\int\limits_{1}^{+\infty} \frac{dx}{x^\alpha} \in \mathbb{R}$;

(b) $\displaystyle\int\limits_{0}^{1} \frac{dx}{x^\alpha} \in \mathbb{R}$;

(c) $\displaystyle\int\limits_{-\infty}^{+\infty} \sin(\alpha x)\,dx \in \mathbb{R}$.

6.D.36. For which $p, q \in \mathbb{R}$ is the integral

$$\int\limits_{2}^{+\infty} \frac{dx}{x^p (\ln x)^q}$$

finite?

6.D.37. Decide if the following is true:

(a) $\displaystyle\int\limits_{-\infty}^{+\infty} \frac{dx}{x^2+3} \in \mathbb{R}$;

(b) $\displaystyle\int\limits_{-\infty}^{+\infty} \frac{dx}{x^2-3} \in \mathbb{R}$;

(c) $\displaystyle\int\limits_{1}^{+\infty} \frac{1+2\sin^3 x}{x^5+x^3+1}\,dx \in \mathbb{R}$.

Solutions of the exercises

6.A.4. $\frac{2\sin x}{\cos^3 x}$.

6.A.5. $p^{(5)}(x) = 12 \cdot 5!$; $p^{(6)}(x) = 0$.

6.A.6. $2^{12}\,e^{2x} + \cos x$.

6.A.7. $f^{(26)}(x) = -\sin x + 2^{26}\,e^{2x}$.

6.A.8. All of them.

6.A.9. (a) $v(0) = 6\,\text{m/s}$; (b) $t = 3\,\text{s}$, $s(3) = 16\,\text{m}$; (c) $v(4) = -2\,\text{m/s}$, $a(4) = -2\,\text{m/s}^2$.

6.A.12. $\frac{\pi}{4} + \frac{1}{2}(x-1) - \frac{1}{4}(x-1)^2 + \frac{1}{12}(x-1)^3$.

6.A.13. (a) $1 + \frac{x^2}{2}$; (b) $1 - \frac{x^2}{2}$; (c) $x - \frac{x^3}{3}$; (d) $x + \frac{x^3}{3}$; (e) $x + x^2 + \frac{x^3}{3}$.

6.A.14. $2(x-1) - (x-1)^2 + \frac{2}{3}(x-1)^3 - \frac{1}{2}(x-1)^4$.

6.A.15. $\frac{-x^3}{3(1+x)^3}$.

6.A.16. $x - \frac{x^3}{6}$; $\sin 1° \approx \frac{\pi}{180} - \frac{\pi^3}{6 \cdot 180^3}$; $\lim_{x \to 0+} \frac{x\,\sin x - x^2}{x^4} = -\frac{1}{6}$.

6.A.17. $\sum_{k=0}^{n} \frac{2^k}{k!}x^k$, $n \geq 8$, $n \in \mathbb{N}$.

6.A.18. $(x-1)^3 + 3(x-1)^2 + (x-1) + 4$.

6.A.26. $1 - \frac{\pi^2}{10^2 \cdot 2} + \frac{\pi^4}{10^4 \cdot 4!}$.

6.A.27. $1 - 3x + \frac{7}{24}x^4$; above the tangent line.

6.A.29. $\left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$.

6.A.30. It's convex on intervals $(-\infty, 0)$ and $(0, 1/2)$; concave on interval $(1/2, +\infty)$. It has only one asymptote, the line $y = \pi/4$ (v $\pm\infty$).

6.A.31. (a) $y = 0$ at $-\infty$; (b) $x = 2$ – horizontal, $y = 1$ v $\pm\infty$.

6.A.32. $y = 0$ for $x \to \pm\infty$.

6.A.33. $y = \ln 10$, $y = x + \ln 3$.

6.B.1. $S_{\Xi n,\,\text{sup}} = \frac{n+1}{n}$, $S_{\Xi n,\,\text{inf}} = \frac{n-1}{n}$; yes, it is.

6.B.15. $2x^3 + 3x^2 - 2x - 13 + \frac{-19x+53}{x^2-2x+4}$.

6.B.16. $x^3 - \frac{1}{3}x + \frac{2}{9} + \frac{5}{9(3x+2)}$.

6.B.17. (a) $\frac{2}{x-2} + \frac{3}{x+2} - \frac{1}{x+3}$; (b) $\frac{2}{x} - \frac{1}{x^2} + \frac{1}{x^2+1} + \frac{x}{(x^2+1)^2}$.

6.B.18. $\frac{5}{x-2} + \frac{3}{x^3} - \frac{3}{x}$.

6.B.19. $\frac{3}{x+1} + \frac{4x-2}{x^2-4x+13}$.

6.B.20. $\frac{1}{x^2} - \frac{2}{x} + \frac{2x-3}{x^2-x+2}$.

6.B.21. $\frac{1}{x} - \frac{1}{x^2} + \frac{1}{x^3} - \frac{1}{x+1}$.

6.B.22. $\frac{A}{x-2} + \frac{B}{x^2} + \frac{C}{x} + \frac{Dx+E}{(3x^2+x+4)^2} + \frac{Fx+G}{3x^2+x+4}$.

6.B.23. $1 + \frac{1}{x^3} - \frac{3}{x} + \frac{5}{x-2}$.

6.B.24. (a) $3\ln|x-2|$; (b) $\frac{1}{(x-2)^2}$.

6.B.44. $\frac{1}{1-a}$ for $a \in (0,1)$, ∞ else.

6.C.2.
$$\sum_{i=0}^{\infty} (-1)^n \frac{2^{2n-1}}{(2n)!} x^{2n},$$
converges for all real x.

6.C.3.
$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{2^{2n-1}}{(2n)!} x^{2n},$$
converges for all real x.

6.C.4.
$$f(x) = \sum_{n=1}^{\infty} \frac{3(-1)^{n+1}}{n} x^n,$$

converges for $x \in (-1, 1]$.

6.C.5. It's good to realize we're expanding $\frac{1}{2}\ln(x)$.
$$f(x) = \sum_{i=0}^{\infty} (-1)^{i+1} \frac{1}{2i}(x-1)^i,$$

Converges on interval $(0, 2]$.

6.C.7. The error belongs to the interval $(0, 1/200)$.

6.C.8. $0 < \int_1^2 \frac{\cos^{10} x}{10} \ln x \, dx < \frac{1}{10}$, $\int_1^2 x \ln x \, dx = \ln 4 - \frac{3}{4}$.

6.C.9. $\int_1^2 \sqrt{x}\, dx = \frac{2}{3}\left(2\sqrt{2}-1\right)$.

6.C.18. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$.

6.C.19. $y = \operatorname{arctg} x$.

6.C.20. Exactly for $x \in \left(-\frac{5}{2}, \frac{5}{2}\right)$, we have
$$\tfrac{1}{5+2x} = \tfrac{1}{5} \sum_{n=0}^{\infty} \left(-\tfrac{2}{5}\right)^n x^n.$$

6.C.21. $\frac{1}{3}\sum_{n=0}^{\infty} \frac{2^n}{3^n} x^n$.

6.C.22.
$$f(x) = 1/2 + \sum_{i=0}^{\infty} \frac{(-1)^{i+1} 2^{2i}}{(2i+1)!}\left(x-\frac{\pi}{4}\right)^{2i+1}.$$

The series converges for all $x \in R$.

6.C.23. $\sum_{n=0}^{\infty} \frac{e}{n!}(x-1)^n$; $\sum_{n=0}^{\infty} \frac{\ln^n 2}{n!} x^n$.

6.C.24. $f(x) = x$, $x \in \mathbb{R}$; yes.

6.C.25. No.

6.C.26. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$.

6.C.27. (a) $1 - \frac{\pi^2}{18^2 \cdot 2!} + \frac{\pi^4}{18^4 \cdot 4!}$; (b) $\frac{1}{2} - \frac{1}{5 \cdot 2^5}$.

6.C.28. $\sum_{n=0}^{\infty} \frac{1}{(2n+1)\, n!} x^{2n+1}$.

6.C.29. $y = \operatorname{arctg} x$.

6.C.30. Exactly for $x \in \left(-\frac{5}{2}, \frac{5}{2}\right)$, we have
$$\tfrac{1}{5+2x} = \tfrac{1}{5} \sum_{n=0}^{\infty} \left(-\tfrac{2}{5}\right)^n x^n.$$

6.C.31. (a) $-\cotg x - x + C$; (b) $\tg x - \cotg x + C$.

6.C.32. (a) $-x^2 \cos x + 2x \sin x + 2\cos x + C$; (b) $e^x\left(x^2 - 2x + 2\right) + C$.

6.C.33. $\frac{x^2}{4}\left(2\ln^2 x - 2\ln x + 1\right) + C$.

6.C.34. $\left(2x - x^2\right)e^x + C$.

6.C.35. (a) $\frac{(2x+5)^{11}}{22} + C$; (b) $-\frac{1}{\ln x} + C$; (c) $-\frac{1}{3}e^{-x^3} + C$; (d) $5\arcsin^3 x + C$; (e) $\frac{\ln^2 x}{2} + C$;
(f) $\operatorname{arctg}^2 \sqrt{x} + C$; (g) $\frac{\sqrt{3}}{3}\operatorname{arctg}\left(\frac{\sqrt{3}}{3}e^x\right) + C$; (h) $2\sin\sqrt{x} - 2\sqrt{x}\cos\sqrt{x} + C$.

6.C.36. For example $1 - x = t^2 x$ gives $\int \frac{-2}{(1+t^2)^4}\, dt$; and $\sqrt{x^2+x+1} = x + y$ leads to $\int \frac{2\, dy}{y^2+2y-2}$.

6.C.37. $\frac{\sqrt{2}}{2}\operatorname{arctg}\left(\sqrt{2}\tg x\right) + C$.

6.C.38. Infinitely many.

6.C.39. For example, f can attain a value of 1 at rational points of the interval and be zero at irrational points.

6.C.40. (a) 2; (b) $\frac{\pi}{4} - \frac{\ln 2}{2}$; (c) $2\ln\left(1+\sqrt{2}\right)$; (d) $2 - \frac{2}{e}$.

6.C.41. $\sqrt{5} - \sqrt{2}$.

6.C.42. $|b| - |a|$.

6.C.43. $\frac{1}{4}\ln 2$.

6.C.44. $\frac{1}{5}\left(e^{\pi} - 2\right)$.

6.C.45. $e - 5e^{-1}$.

6.C.46. $\frac{1}{6}$.

6.C.47. (a) 4; (b) $\frac{1 - \ln 2}{2}$.

6.C.48. $p < q$.

6.C.49. $a > 0$; $b = 0$; $c > 0$.

6.C.50. $C < D = 0 < A < B$.

6.C.51. (a) 5; (b) 0; (c) 0.

6.C.52. 1.

6.C.53. 0.

6.C.54. $-6x^5 \cos x^2$.

6.D.1. The range is $(-\infty, 0] \cup [4, +\infty)$. Function f is not odd, even nor periodic. It has a single discontinuity $x_0 = -1$ with
$$\lim_{x \to -1+} f(x) = -\infty, \qquad \lim_{x \to -1-} f(x) = +\infty.$$
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
$$\lim_{x \to -\infty} f(x) = +\infty, \qquad \lim_{x \to +\infty} f(x) = -\infty;$$
$$f'(x) = -\frac{x^2 + 2x}{(x+1)^2}, \quad f''(x) = -\frac{2}{(x+1)^3}, \quad x \in \mathbb{R} \setminus \{-1\}.$$
This implies that f is increasing on the intervals $[-2, -1)$, $(-1, 0]$ and decreasing on the intervals $(-\infty, -2]$, $[0, +\infty)$. At the stationary point $x_1 = 0$ it reaches a strict local maximum and at the stationary point $x_2 = -2$ it has a local minimum $y_2 = 4$. It is convex on the interval $(-\infty, -1)$ and concave on the interval $(-1, +\infty)$. It does not have a point of inflection. The line $x = -1$ is a horizontal asymptote, the inclined asymptote at $\pm\infty$ is the line $y = -x + 1$. For example, $f(-3) = 9/2$, $f'(-3) = -3/4$, $f(1) = -1/2$, $f'(1) = -3/4$.

6.D.1. The function is defined and continuous on $\mathbb{R} \setminus \{0\}$. It is not odd, even, nor periodic. It is negative on the interval $(1, +\infty)$. The only point of intersection of the graph with the axes is the point $[1, 0]$. At the origin, f has a discontinuity of the second kind and its range is is $\mathbb{R}$, because
$$\lim_{x \to 0} f(x) = +\infty, \qquad \lim_{x \to +\infty} f(x) = -\infty, \qquad \lim_{x \to -\infty} f(x) = +\infty.$$
Moreover,
$$f'(x) = -\frac{x^3 + 2}{x^3}, \quad x \in \mathbb{R} \setminus \{0\},$$
$$f''(x) = \frac{6}{x^4}, \quad x \in \mathbb{R} \setminus \{0\}.$$
The only stationary point is $x_1 = -\sqrt[3]{2}$. The function f is increasing on the interval $[x_1, 0)$, decreasing on the intervals $(-\infty, x_1]$, $(0, +\infty)$. Hence at x_1 it has a local minimum $y_1 = 3/\sqrt[3]{4}$. It has no points of inflection. It is convex on its whole domain. The line $x = 0$ is a horizontal asymptote and the line $y = -x$ is an inclined asymptote at $\pm\infty$.

6.D.3. The function is defined and continuous on $\mathbb{R} \setminus \{1\}$. It is not odd, even nor periodic. The points of intersection of the graph of f with the axes are the points $\left[1 - \sqrt[3]{2}, 0\right]$ 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
$$\lim_{x \to 1-} f(x) = -\infty, \qquad \lim_{x \to 1+} f(x) = +\infty, \qquad \lim_{x \to \pm\infty} f(x) = +\infty.$$
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 $\left(-\infty, 1 - \sqrt[3]{2}\right)$, $(1, +\infty)$ and concave on the intervals $\left(1 - \sqrt[3]{2}, 1\right)$. 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.4. The function is defined and continuous on $\mathbb{R}$. 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:

$$f'(x) = \frac{\mathrm{e}^{-x}}{3\sqrt[3]{x^2}} - \sqrt[3]{x}\,\mathrm{e}^{-x}, \quad x \in \mathbb{R} \smallsetminus \{0\}, \quad f'(0) = +\infty,$$

$$f''(x) = \sqrt[3]{x}\,\mathrm{e}^{-x} - \frac{2\mathrm{e}^{-x}}{3\sqrt[3]{x^2}} - \frac{2\mathrm{e}^{-x}}{9\sqrt[3]{x^5}}, \quad x \in \mathbb{R} \smallsetminus \{0\}.$$

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 $y_0 = 1/\sqrt[3]{3\mathrm{e}}$. Since $\lim_{x \to -\infty} f(x) = -\infty$, its range is $(-\infty, y_0]$. The points of inflection are

$$x_1 = \tfrac{1-\sqrt{3}}{3}, \quad x_2 = 0, \quad x_3 = \tfrac{1+\sqrt{3}}{3}.$$

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. $\lim_{x \to +\infty} f(x) = 0$.

6.D.5. The function is defined and continuous on $\mathbb{R} \smallsetminus \{2\}$. It is not odd, even nor periodic. It's positive exactly on the interval $(0, 2)$. The only point of intersection of the graph of f with the axes is the point $[0, 0]$. At $x_0 = 2$, the jump of size π is observed, as follows from the limits

$$\lim_{x \to 2-} f(x) = \tfrac{\pi}{2}, \quad \lim_{x \to 2+} f(x) = -\tfrac{\pi}{2}.$$

We have

$$f'(x) = \tfrac{1}{x^2 - 2x + 2}, \quad x \in \mathbb{R} \smallsetminus \{2\},$$

$$f''(x) = \tfrac{2(1-x)}{(x^2 - 2x + 2)^2}, \quad x \in \mathbb{R} \smallsetminus \{2\}.$$

The first derivative does not have a zero point. The function f is therefore increasing at every point of its domain. Since

$$\lim_{x \to -\infty} f(x) = -\tfrac{\pi}{4}, \quad \lim_{x \to +\infty} f(x) = -\tfrac{\pi}{4},$$

its range is the set $(-\pi/2, \pi/2) \smallsetminus \{-\pi/4\}$. The function f is convex on the interval $(-\infty, 1)$, concave on the intervals $(1, 2)$, $(2, +\infty)$. Thus the point $x_1 = 1$ is a point of inflection with $f(1) = \pi/4$. The only asymptote is the line $y = -\pi/4$ at $\pm\infty$.

6.D.6. Domain $\mathbb{R}^+$, global maximum $x = e$, point of inflection

$$x = \sqrt{e^3}$$

, increasing on $(0, e)$, decreasing on (e, ∞), concave on $(0, \sqrt{e^3})$, convex on $(\sqrt{e^3}, \infty)$, asymptotes $x = 0$ and $y = 0$, $\lim_{x \to 0} f(x) = -\infty$, $\lim_{x \to \infty} f(x) = 0$.

6.D.7. Domain $\mathbb{R} \setminus [1, 2]$. Local maximum $x = \tfrac{1-\sqrt{5}}{2}$, concave on the whole domain, asymptotes $x = 1$, $x = 2$.

6.D.8. Domain $\mathbb{R}$. Local minimas $-1, 1$, maximum at 0. Even function. Points of inflection $\pm\tfrac{1}{\sqrt{2}}$, no asymptotes.

6.D.9. Domain $\mathbb{R} \setminus [-\tfrac{1}{2}, 1]$. No global extremes. No inflection points, asymptotes $x = -\tfrac{1}{2}$, $x = 1$.

6.D.10. Domain $\mathbb{R} \setminus \{1\}$. No extremes. No points of inflection, convex on $(-\infty, 1)$, concave on $(1, \infty)$. Horizontal asymptote $x = 1$. Inclined asymptote $y = x + 1$.

6.D.11. (a) $\tfrac{8}{15} x\sqrt[8]{x^7}$; (b) $\tfrac{4^x}{\ln 4} + 2\tfrac{6^x}{\ln 6} + \tfrac{9^x}{\ln 9}$; (c) $\tfrac{\arcsin x}{2}$; (d) $\ln(1 + \sin x)$.

6.D.12. $\mathrm{e}^x + 3 \arcsin \tfrac{x}{2}$.

6.D.13. $\tfrac{1}{4} \ln(1 + x^4) + C$.

6.D.14. $2\sqrt{2} \operatorname{arctg} \tfrac{x-1}{\sqrt{2}} + C$.

6.D.15. $\ln\left|\tfrac{x^2-1}{x}\right| + \tfrac{3}{2}\arcsin x - 4\cos x - 5\sin x + C$.

6.D.16. (a) $x \operatorname{arctg} x - \tfrac{\ln(1+x^2)}{2} + C$; (b) $\tfrac{\ln^2 x}{2} + C$.

6.D.17. $x - 2\sqrt{x} + 2\ln(1 + \sqrt{x}) + C$.

6.D.18. (a) $\tfrac{x^{n+1}}{n+1} \ln x - \tfrac{x^{n+1}}{(n+1)^2} + C$; (b) $\tfrac{\operatorname{arctg} x^2}{2} + C$.

6.D.19. $\tfrac{3}{2} \ln(x^2 + 4x + 8) - \tfrac{1}{2}\operatorname{arctg}\tfrac{x+2}{2} + C$.

6.D.20. $\tfrac{4}{3\sqrt{3}}\operatorname{arctg}\tfrac{2x+1}{\sqrt{3}} + \tfrac{2x+1}{3(x^2+x+1)} + C$.

6.D.21. $\tfrac{1}{6}\ln\tfrac{(x+1)^2}{x^2-x+1} + \tfrac{\sqrt{3}}{3}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} + C$.

6.D.22. $\tfrac{1}{3}\ln|x-1| - \tfrac{1}{6}\ln(x^2 + x + 1) - \tfrac{1}{\sqrt{3}}\operatorname{arctg}\tfrac{2x+1}{\sqrt{3}} + C$.

6.D.23. $\ln\left(|x-1|(x-2)^4\right) - \frac{8}{x-2} + x + C.$

6.D.24. (a) $\frac{\cos^7 x}{7} - \frac{\cos^5 x}{5} + C$; (b) $\frac{\operatorname{tg} x}{2} + \frac{x}{2} + C$; (c) $x - \sin x + C$; (d) $\frac{x}{2} + \frac{\sin 2x}{4} + C$; (e) $\frac{2}{3}\sin^{\frac{3}{2}} x - \frac{4}{7}\sin^{\frac{7}{2}} x + \frac{2}{11}\sin^{\frac{11}{2}} x + C$; (f) $\frac{\operatorname{tg}^3 x}{3} + 2\operatorname{tg} x - \frac{1}{\operatorname{tg} x} + C$; (g) $\frac{1}{2}\ln\left|\operatorname{tg}\frac{x}{2}\right| - \frac{\cos x}{2\sin^2 x} + C$; (h) $\ln\left|\operatorname{tg}\frac{x}{2}\right| + C.$

6.D.25. $\frac{1}{2}\ln(x^2 + 2x + 2) - \frac{1}{2}\ln(x^2 + x + 1) + \frac{1}{3}\sqrt{3}\arctan\left(\frac{(2x+1)\sqrt{3}}{3}\right) + C.$

6.D.26. $\frac{1}{2}\ln\left(\frac{2+\ln 2}{2-\ln 2}\right).$

6.D.27. $-\frac{1}{6} - \frac{2}{9}\ln 2.$

6.D.28.

(i) $\frac{2}{3}$,

(ii) $\frac{1}{2}\sin^4 x.$

6.D.29. (a) π; (b) $+\infty$; (c) 20; (d) -2.

6.D.30. $-\infty$.

6.D.31. $\frac{\sqrt{3}}{9}\pi.$

6.D.32. $\frac{2\sqrt{3}}{9}\pi.$

6.D.33. $-\frac{1}{2}$; 1.

6.D.34. $2 - \frac{2}{e}$; $\frac{2}{e} - \frac{2}{e^2}$; $\frac{2}{e^2}$.

6.D.35. (a) $\alpha > 1$; (b) $\alpha < 1$; (c) $\alpha = 0$.

6.D.36. For $p > 1$, $q \in \mathbb{R}$ and for $p = 1$, $q > 1$.

6.D.37. (a) true; (b) false; (c) true.

Continuous tools for modelling

How do we manage non-linear objects?
– mainly by linear tools again...

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.

A. Orthogonal systems of functions

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. Perhaps we recall from Chapter 2 that the orthogonal projections

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).

were easily computed in terms of inner products. Now we do the same for the infinite dimensional spaces of functions.

The inner product mimics the product of scalars. Actually, the simplest way is to introduce the inner product to suitable vector spaces of functions on a given interval $I = [a, b] \subset \mathbb{R}$ in this way:

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,\mathrm{d}\,x.$$

We refer to this inner product as L_2, see 7.1.3 for more information, including the complex valued functions.

Such a scalar product allows us to calculate the projections to finite dimensional subspaces in the same way as we did in finite dimensional vector spaces.

7.A.1. Given the vector subspace $\langle x^2, 1/x \rangle$ of the space of real-valued functions on the interval $[1, 2]$, with the L_2 product, complete the function $1/x$ to an orthogonal basis of the subspace. Determine the orthogonal projection of the function x onto it, and compute the distance of the function x from this subspace.

Solution. First, consider the basis. It is required that the function $1/x$ be one of the vectors of the basis. The vector space in question is generated by two linearly independent functions, thus its dimension is 2. All of the vectors in it are of the form $a \cdot \frac{1}{x} + b \cdot x^2$ for some $a, b \in \mathbb{R}$. It remains to find one more vector of the basis which is orthogonal to the function $f_1 = 1/x$. According to the Gram–Schmidt process, we seek it in the form $f_2 = x^2 + k \cdot \frac{1}{x}$, $k \in \mathbb{R}$. The real constant k can be determined from the condition of orthogonality:

$$0 = \left\langle \frac{1}{x}, x^2 + k \cdot \frac{1}{x} \right\rangle = \left\langle \frac{1}{x}, x^2 \right\rangle + k \left\langle \frac{1}{x}, \frac{1}{x} \right\rangle.$$

Therefore,

$$k = -\frac{\langle \frac{1}{x}, x^2 \rangle}{\langle \frac{1}{x}, \frac{1}{x} \rangle} = -\frac{\int_1^2 \frac{1}{x} \cdot x^2 \,\mathrm{d}x}{\int_1^2 \frac{1}{x} \cdot \frac{1}{x} \,\mathrm{d}x} = -3.$$

Thus, the requested orthogonal basis is $(\frac{1}{x}, x^2 - \frac{3}{x})$.

Next, we calculate the projection p_x of the function x onto this subspace (see (1) on page 110). We find

$$p_x = \frac{\langle x, \frac{1}{x} \rangle}{\langle \frac{1}{x}, \frac{1}{x} \rangle} \cdot \frac{1}{x} + \frac{\langle x, x^2 - \frac{3}{x} \rangle}{\langle x^2 - \frac{3}{x}, x^2 - \frac{3}{x} \rangle} \cdot \left(x^2 - \frac{3}{x} \right)$$
$$= \frac{2}{x} + \frac{15}{34}\left(x^2 - \frac{3}{x} \right).$$

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 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 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 (ie. 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

Finally, 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:

$$\|x - p_x\|_2 \;=\; \left(\int_1^2 \left(x - \frac{2}{x} - \frac{15}{34}\left(x^2 - \frac{3}{x}\right)\right)^2 dx \right)^{\frac{1}{2}}$$

$$=\; \frac{1}{\sqrt{408}} \doteq 0.0495. \qquad \square$$

7.A.2. Consider the real vector space of functions on the interval $[1,2]$ generated by the functions $\frac{1}{x}, \frac{1}{x^2}, \frac{1}{x^3}$ with the L_2 product. Complete the function $\frac{1}{x}$ to an orthogonal basis of this space. Determine also the projection of the functions $\frac{1}{x^4}$ and x onto the above vector space. Then find their distances from this vector space.

Solution. As in the previous exercise, use the Gram–Schmidt orthogonalization process (with the given scalar product). Successively,

$$f_1(x) = \frac{1}{x},$$
$$f_2(x) = \frac{1}{x^2} - \frac{3}{4x},$$
$$f_3(x) = \frac{1}{x^3} - \frac{3}{2x^2} + \frac{13}{24x}.$$

The projection of $\frac{1}{x^4}$ is $\frac{15}{32}f_1 + \frac{69}{40}f_2 + \frac{9}{4}f_3$ while the distance is $\frac{\sqrt{14}}{2240} = 0.00167$.

The projection of x is $2f_1 + 96(-\frac{3}{4} + \ln(2))f_2 + 5760(-\frac{3}{2}\ln(2) + \frac{25}{24})f_3$, while the distance is approximately 0.035.
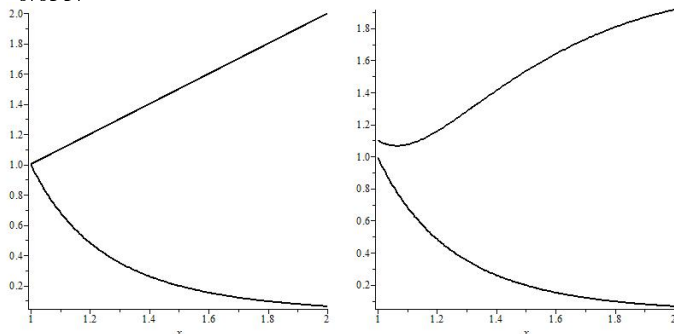
The illustrations show the functions x and $1/x^4$ and their approximations. We can see that the function $\frac{1}{x^4}$, which is the one whose shape is similar to that of one or more generators, is approximated much better by the projection. $\qquad \square$

But there are plenty of inner products on functions. We mention two of them in the following exercises.

7.A.3. Verify that for each interval $I = [a,b] \subset \mathbb{R}$ and positive continuous function ω on I, the formula

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,

$$\int_a^b |f(x) - g(x)|\, dx = \int_a^b |f(x) - h(x) + h(x) - g(x)|\, dx$$

$$\leq \int_a^b |f(x) - h(x)|\, dx + \int_a^b |h(x) - g(x)|\, dx,$$

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

$$\langle f, g \rangle_\omega = \int_a^b f(x)g(x)\omega(x)\,\mathrm{d}\,x$$

defines an inner product on the continuous functions on I. Verify that choosing $I = (-1, 1)$ and $\omega(x) = (1 - x^2)^{-1/2}$, the functions

$$T_k(x) = \cos(k \arccos(x)), \quad k \in \mathbb{N},$$

form an orthogonal system of polynomials with respect to this inner product. These polynomials are called *Chebyshev polynomials*.

Solution. We compare the defining formula with the L_2 inner product above: Consider the substitution $x = \varphi(z)$, where φ is the inverse function to $z = \int_a^x \omega(t)\,dt$. The inverse exists, since ω is positive and so z is a strictly increasing function of x. Thus, $dz = \omega(x)dx$ and

$$
\begin{aligned}
\langle f, g \rangle_\omega &= \int_a^b f(x)g(x)\omega(x)\,\mathrm{d}\,x \\
&= \int_0^{\varphi^{-1}(b)} f(\varphi(z))g(\varphi(z))\,\mathrm{d}\,z.
\end{aligned}
$$

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 $\langle\ ,\ \rangle_\omega$ is an inner product. Of course, the properties of the inner product can be checked directly.

In this special case, $\omega(x) = \frac{\mathrm{d}}{\mathrm{d}\,x}(\arccos(x))$, and thus the above substitution yields

$$\langle T_r, T_s \rangle_\omega = \int_0^\pi \cos(rz)\cos(sz)\,\mathrm{d}\,z.$$

We are dealing with improper Riemann integrals (integrating the unbounded function ω), but this does not cause any problem. By using the well known trigonometric formula

$$\cos(rz)\cos(sz) = \frac{1}{2}\big(\cos((r - s)z) + \cos((r + s)z)\big)$$

the integral is easily evaluated. It vanishes for all $r \neq s$.

The remaining task is to show that $T_k(x)$ is a polynomial. The definition itself shows directly that

$$T_0(x) = 1, \quad T_1(x) = x.$$

By the above trigonometric formula,

$$
\begin{aligned}
T_{n+1}&(x) + T_{n-1}(x) \\
&= \cos\big((n + 1)\arccos(x)\big) + \cos\big((n - 1)\arccos(x)\big) \\
&= 2\cos\big(n\arccos(x)\big)\cos\big(\arccos(x)\big) = 2xT_n(x)
\end{aligned}
$$

so that for all positive n,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

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.

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 $f_1, \ldots, f_k$ (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 $f_i(x) = x^i$, $i = 0, \ldots, k$. 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 $f_0, \ldots, f_k$. Moreover, if we can arrange the choice of generators to satisfy

$$(1) \qquad \langle f_i, f_j \rangle = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases}$$

then the computations become much easier than they would otherwise.

Recall the Gram–Schmidt orthogonalization procedure, see 2.3.20. 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

$$g_{\ell+1} = f_{\ell+1} + a_1 g_1 + \cdots + a_\ell g_\ell, \quad a_i = -\frac{\langle f_{\ell+1}, g_i \rangle}{\|g_i\|^2}$$

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

$$g_1 = 1$$

$$g_2 = x - \frac{1}{\|g_1\|^2}\left(\int_{-1}^1 x \cdot 1\,dx\right) \cdot g_1 = x - 0 = x$$

$$g_3 = x^2 - \frac{1}{\|g_1\|^2}\left(\int_{-1}^1 x^2 \cdot 1\,dx\right) \cdot g_1 -$$

$$\frac{1}{\|g_2\|^2}\left(\int_{-1}^1 x^2 \cdot x\,dx\right) \cdot g_2 = x^2 - \frac{1}{3}.$$

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

This is the recurrent definition of Chebyshev polynomials. That all $T_k(x)$ are polynomials now follows by induction. $\square$

7.A.4. Show that the choice of the weight function $\omega(x) = e^{-x}$ and the interval $I = [0, \infty)$ in the previous example 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. $\bigcirc$

7.A.5. 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 $1, x, x^2, \ldots, x^n, \ldots$, using the inner products $\langle\ ,\ \rangle_\omega$, possibly only up to signs. $\bigcirc$

Given a finite-dimensional vector (sub)space of functions, calculate first the orthogonal (or orthonormal) basis of this subspace by the Gram–Schmidt orthogonalization process (see 2.3.20). Then determine the orthogonal projection as before. See the formula (1) at page 110.
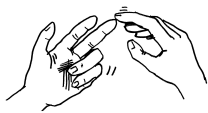
7.A.6. Given the vector subspace $\langle \sin(x), x \rangle$ of the space of real-valued functions on the interval $[0, \pi]$, complete the function x to an orthogonal basis of the subspace and determine the orthogonal projection of the function $\frac{1}{2}\sin(x)$ onto it. $\bigcirc$

7.A.7. Given the vector subspace $\langle \cos(x), x \rangle$ of the space of real-valued functions 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. $\bigcirc$

B. Fourier series

Having a countable system of orthogonal functions T_k, $k = 0, 1, \ldots$, as in the examples above, we may sequentially project a given function f to the subspaces $V_k = \langle T_0, \ldots, T_k \rangle$. If the limit of these projections exists, this determines a series built on linear combinations of T_k. Under additional conditions, this should allow us to differentiate or integrate the function f in a similar way, as we did with the power series.

We consider one particular orthogonal system of periodic functions, namely that of J.B.J. Fourier.

The periodic functions are those describing periodic processes, i.e. $f(t+T) = f(t)$ for some positive constant $T \in \mathbb{R}$,

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, $i = 0, \ldots$, 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*. If all the functions f_n in the sequence are pairwise orthogonal, and if for all n, the norm $\|f_n\|_2 = 1$, we talk about an *orthonormal system of functions*.

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.

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

$$\langle F, f_n \rangle = \sum_{m=0}^{\infty} c_m \int_a^b f_m(x)\overline{f_n(x)}\, dx = c_n \|f_n\|_2{}^2,$$

called the period of f, and all $t \in \mathbb{R}$. One of the fundamental periodic processes which occur in applications is a general simple harmonic oscillation in mechanics. The function $f(t)$ which describes the position of the point mass on the line in the time t is of the form
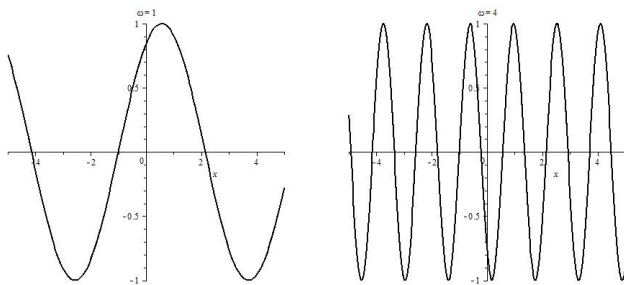
$$(1) \qquad f(t) = a \sin(\omega t + b)$$

for certain constants $a, \omega > 0$, $b \in \mathbb{R}$. In the diagram on the left, $f(t) = \sin(t + 1)$ and on the right, $f(t) = \sin(4t + 4)$:

Applying the standard trigonometric formula

$$\sin(\alpha + \beta) = \cos \alpha \sin \beta + \sin \alpha \cos \beta,$$

with $\alpha, \beta \in \mathbb{R}$, we write the function $f(t)$ alternatively as

$$(2) \qquad f(t) = c \cos(\omega t) + d \sin(\omega t),$$

where $c = a \sin b$, $d = a \cos b$.

7.B.1. Show that the system of functions 1, $\sin(x)$, $\cos(x)$, $\ldots$, $\sin(nx)$, $\cos(nx)$, $\ldots$ is orthogonal with respect to the L_2 inner product on the interval $I = [-\pi, \pi]$. $\bigcirc$

Building an orthogonal system of periodic functions $\sin(nx)$ and $\sin(nx + \pi/2) = \cos(nx)$ leads to the classical *Fourier series*. [1]

In application problems, we often meet the superposition of different harmonic oscillations. The superposition of finitely many harmonic oscillations is expressed by sums of functions of the form

$$f_n(x) = a_n \cos(n\omega x) + b_n \sin(n\omega x)$$

for $n \in \{0, 1, \ldots, m\}$. These particular functions have prime period $2\pi/(n\omega)$. Therefore, their sum

$$(3) \qquad \frac{a_0}{2} + \sum_{n=1}^{m} \big(a_n \cos(n\omega x) + b_n \sin(n\omega x)\big)$$

is a periodic function with period $T = 2\pi/\omega$.

---

[1] The Fourier series are named in honour of the French mathematician and physicist Jean B. J. Fourier, who was the first to apply the Fourier series in practice in his work from 1822 devoted to the issue of heat conduction (he began to deal with this issue in 1804-1811). He introduced mathematical methods which even nowadays lie at the core of theoretical physics. He did not pay much attention to physics himself.

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.3 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, $n = 1, 2, \ldots$, 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 $f_1, \ldots, f_n$ 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 $\lim_{n \to \infty} \|g - h_n\|_2 = 0$.

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 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.

7.B.2. Show that the system of functions 1, $\sin(n\omega x)$, $\cos(n\omega x)$, for all positive integers n is orthogonal with respect to the L_2 inner product on the interval $[-\pi/\omega, \pi/\omega]$.
○

When projecting a given function orthogonally to the subspace of functions (3), the key concept is the set of *Fourier coefficients* a_n and b_n, $n \in \mathbb{N}$.

7.B.3. Find the Fourier series for the periodic extension of the function

(a) $g(x) = 0$, $x \in [-\pi, 0)$, $\quad g(x) = \sin x$, $x \in [0, \pi)$;
(b) $g(x) = |x|$, $x \in [-\pi, \pi)$;
(c) $g(x) = 0$, $x \in [-1, 0)$, $\quad g(x) = x + 1$, $x \in [0, 1)$.

Solution. Before starting the computations, consider the illustrations of the resulting approximation of the given functions. The first two display the finite approximation in the cases a) and b) up to $n = 5$, while the third illustration for the case c) goes up to $n = 20$. Clearly the approximation of the discontinuous function is much slower and it also demonstrates the Gibbs phenomenon. This is the overshooting in the jumps, which is proportional to the magnitudes of the jumps.
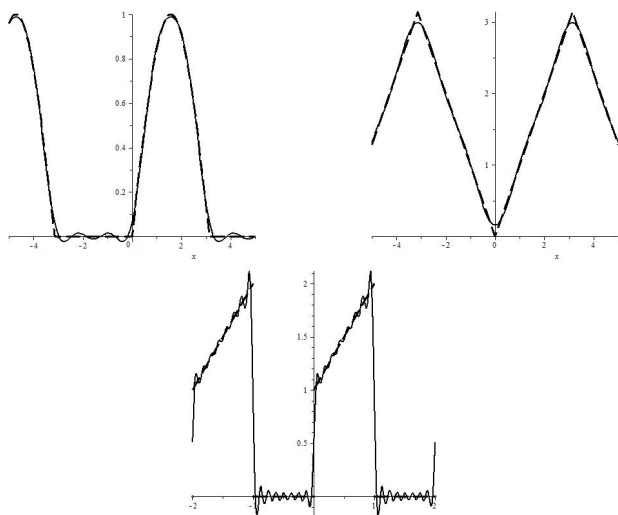
The case (a). Direct calculation gives (using formulae from 7.1.6)

$$a_0 = \frac{1}{\pi} \int\limits_{-\pi}^{\pi} g(x)\,\mathrm{d}x = \frac{1}{\pi} \int\limits_{-\pi}^{0} 0\,\mathrm{d}x + \frac{1}{\pi} \int\limits_{0}^{\pi} \sin x\,\mathrm{d}x$$
$$= \frac{1}{\pi}\left[-\cos x\right]_0^{\pi} = \frac{2}{\pi},$$

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 any given 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

$$\left\| g - \sum_{n=1}^{k} a_n f_n \right\|^2 = \int_a^b \left| g(x) - \sum_{n=1}^{k} a_n f_n(x) \right|^2 dx$$

$$= \int_a^b |g(x)|^2\,dx - \int_a^b \sum_{n=1}^{k} g(x)\overline{a_n f_n(x)}\,dx -$$

$$- \int_a^b \sum_{n=1}^{k} a_n f_n(x)\overline{g(x)}\,dx + \int_a^b \left| \sum_{n=1}^{k} a_n f_n(x) \right|^2 dx$$

$$= \|g\|^2 - \sum_{n=1}^{k} \overline{a_n} c_n \|f_n\|^2 - \sum_{n=1}^{k} a_n \overline{c_n}\|f_n\|^2 + \sum_{n=1}^{k} a_n^2 \|f_n\|^2$$

$$= \|g\|^2 + \sum_{n=1}^{k} \|f_n\|^2\big((c_n - a_n)\overline{(c_n - a_n)} - |c_n|^2\big)$$

$$= \|g\|^2 + \sum_{n=1}^{k} \|f_n\|^2\big(|c_n - a_n|^2 - |c_n|^2\big).$$

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 completes the proof of the first statement. With this choice of a_n, we obtain *Bessel's identity*

$$\left\| g - \sum_{n=1}^{k} c_n f_n \right\|^2 = \|g\|^2 - \sum_{n=1}^{k} |c_n|^2 \|f_n\|^2.$$

Since the left-hand side is non-negative, it follows that:

$$\sum_{n=1}^{k} c_n^2 \|f_n\|^2 \leq \|g\|^2.$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \cos(nx) \, \mathrm{d}x$$

$$= \frac{1}{\pi} \int_{-\pi}^{0} 0 \, \mathrm{d}x + \frac{1}{\pi} \int_{0}^{\pi} \sin x \cos(nx) \, \mathrm{d}x$$

$$= \frac{1}{2\pi} \int_{0}^{\pi} \sin((1+n)x) + \sin((1-n)x) \, \mathrm{d}x$$

$$= \frac{1}{2\pi} \left[-\frac{\cos((1+n)x)}{1+n} - \frac{\cos((1-n)x)}{1-n} \right]_{0}^{\pi}$$

$$= \frac{1}{2\pi} \left(-\frac{\cos((1+n)\pi)}{1+n} - \frac{\cos((1-n)\pi)}{1-n} + \frac{1}{1+n} + \frac{1}{1-n} \right)$$

$$= \frac{1}{2\pi} \left(-\frac{(-1)^{1+n}}{1+n} - \frac{(-1)^{1-n}}{1-n} + \frac{1}{1+n} + \frac{1}{1-n} \right)$$

$$= \frac{1}{2\pi} \left(\frac{(-1)^{n}+1}{1+n} + \frac{(-1)^{n}+1}{1-n} \right)$$

$$= \frac{1}{\pi} \left(\frac{(-1)^{n}+1}{1-n^2} \right), \quad n \in \mathbb{N}, \ n \neq 1,$$

$$a_1 = \frac{1}{2\pi} \int_{0}^{\pi} \sin(2x) \, \mathrm{d}x = 0$$

$$b_1 = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \sin x \, \mathrm{d}x = \frac{1}{\pi} \int_{-\pi}^{0} 0 \, \mathrm{d}x + \frac{1}{\pi} \int_{0}^{\pi} \sin^2 x \, \mathrm{d}x$$

$$= \frac{1}{2\pi} \int_{0}^{\pi} 1 - \cos(2x) \, \mathrm{d}x = \frac{1}{2\pi} \left[x - \frac{\sin(2x)}{2} \right]_{0}^{\pi} = \frac{1}{2},$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \sin(nx) \, \mathrm{d}x$$

$$= \frac{1}{\pi} \int_{-\pi}^{0} 0 \, \mathrm{d}x + \frac{1}{\pi} \int_{0}^{\pi} \sin x \sin(nx) \, \mathrm{d}x$$

$$= \frac{1}{2\pi} \int_{0}^{\pi} \cos((1-n)x) - \cos((1+n)x) \, \mathrm{d}x$$

$$= \frac{1}{2\pi} \left[\frac{\sin((1-n)x)}{1-n} - \frac{\sin((1+n)x)}{1+n} \right]_{0}^{\pi} = 0,$$

for $n \in \mathbb{N} \smallsetminus \{1\}$.

Thus, we arrive at the Fourier series

$$\frac{1}{\pi} + \frac{\sin x}{2} + \frac{1}{\pi} \sum_{n=2}^{\infty} \left(\frac{(-1)^{n}+1}{1-n^2} \cos(nx) \right).$$

Since $(-1)^n + 1 = 0$ when n is odd, and $= 2$ when n is even, we can put $n = 2m$ to obtain the series

$$\frac{1}{\pi} + \frac{\sin x}{2} - \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{\cos(2mx)}{4m^2 - 1}.$$

The case (b). The given function is of a sawtooth-shaped oscillation. Its expression as a Fourier series is very important in practice. Since the function g is even on $(-\pi, \pi)$, it is immediate that $b_n = 0$ for all $n \in \mathbb{N}$. Therefore, it suffices to determine a_n for $n \in \mathbb{N}$:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \, \mathrm{d}x = \frac{2}{\pi} \int_{0}^{\pi} x \, \mathrm{d}x = \frac{2}{\pi} \left[\frac{x^2}{2} \right]_{0}^{\pi} = \pi.$$

For other $n \in \mathbb{N}$, use integration by parts, to get

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \cos(nx) \, \mathrm{d}x = \frac{2}{\pi} \int_{0}^{\pi} x \cos(nx) \, \mathrm{d}x$$

$$= \frac{2}{\pi} \left[\frac{x}{n} \sin(nx) \right]_{0}^{\pi} - \frac{2}{n\pi} \int_{0}^{\pi} \sin(nx) \, \mathrm{d}x$$

$$= \frac{2}{n^2 \pi} \left[\cos(nx) \right]_{0}^{\pi} = \frac{2}{n^2 \pi} \left((-1)^{n} - 1 \right).$$

So $a_n = -\frac{4}{n^2 \pi}$ for n odd, $a_n = 0$ for n even.

Let $k \longrightarrow \infty$. 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. $\square$

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

$$F(x) = \sum_{n=1}^{\infty} c_n f_n$$

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

$$1, \ \sin x, \ \cos x, \ \sin 2x, \ \cos 2x, \ \ldots, \ \sin nx, \ \cos nx, \ \ldots$$

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 2π (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".

This determines the Fourier series of a function of sawtooth-shaped oscillation as

$$\frac{\pi}{2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \left(\frac{(-1)^n - 1}{n^2} \cos(nx) \right)$$

$$= \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2}$$

$$= \frac{\pi}{2} - \frac{4}{\pi} \left(\cos x + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \cdots \right).$$

This series could have been found by an easier means, namely by integrating the Fourier series of Heaviside's function (see "square wave function" in 7.1.9).

The case (c). The period for this function is $T = 2$, and so $\omega = 2\pi/T = \pi$. Use the more general formulae from 7.1.6, namely

$$a_0 = \frac{2}{T} \int_{x_0}^{x_0+T} g(x)\, \mathrm{d}x = \int_{-1}^{1} g(x)\, \mathrm{d}x$$
$$= \int_{-1}^{0} 0\, \mathrm{d}x + \int_{0}^{1} (x+1)\, \mathrm{d}x = \frac{3}{2},$$

$$a_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \cos(n\omega x)\, \mathrm{d}x = \int_{-1}^{1} g(x) \cos(n\pi x)\, \mathrm{d}x$$
$$= \int_{-1}^{0} 0\, \mathrm{d}x + \int_{0}^{1} (x+1) \cos(n\pi x)\, \mathrm{d}x = \frac{(-1)^n - 1}{n^2\pi^2},$$

$$b_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \sin(n\omega x)\, \mathrm{d}x = \int_{-1}^{1} g(x) \sin(n\pi x)\, \mathrm{d}x$$
$$= \int_{-1}^{0} 0\, \mathrm{d}x + \int_{0}^{1} (x+1) \sin(n\pi x)\, \mathrm{d}x = \frac{1 - 2(-1)^n}{n\pi}.$$

The calculation of a_0 was simple and needs no further comment. As for determining the integrals at a_n and b_n, it sufficed to use integration by parts once. Thus, the desired Fourier series is

$$\frac{3}{4} + \sum_{n=1}^{\infty} \left(\frac{(-1)^n - 1}{n^2\pi^2} \cos(n\pi x) + \frac{1 - 2(-1)^n}{n\pi} \sin(n\pi x) \right).$$

Some refinements of the expression are available. For instance, for $n \in \mathbb{N}$,

$$a_n = -\frac{2}{n^2\pi^2} \text{ for } n \text{ odd}, \quad a_n = 0 \text{ for } n \text{ even},$$

and, similarly,

$$b_n = \frac{3}{n\pi} \text{ for } n \text{ odd}, \quad b_n = -\frac{1}{n\pi} \text{ for } n \text{ even}.$$

$\square$

In the next exercise we show that the calculation of the Fourier series does not always require an integration. Especially in the case when the function g is a sum of products (powers) of functions $y = \sin(mx)$, $y = \cos(nx)$ for $m, n \in \mathbb{N}$, one can rewrite g as a finite linear combination of basic functions.

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 $0 \le x_0 < T$, using a translation by a suitable multiple of T. Then,

$$\int_{x_0}^{x_0+T} f(x)\, dx = \int_{x_0}^{T} f(x)\, dx + \int_{T}^{x_0+T} f(x)\, dx$$
$$= \int_{x_0}^{T} f(x)\, dx + \int_{0}^{x_0} f(x)\, dx = \int_{0}^{T} f(x)\, dx$$

FOURIER SERIES

The series of functions

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(nx) + b_n \sin(nx) \right)$$

from the theorem 7.1.5, 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*.

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 2π. 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 $[x_0, x_0 + T]$:

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(n\omega x) + b_n \sin(n\omega x) \right),$$

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.B.4. Determine the Fourier coefficients of the function

(a) $g(x) = \sin(2x)\cos(3x), \ x \in [-\pi, \pi]$;

(b) $g(x) = \cos^4 x, \ x \in [-\pi, \pi]$.

Solution. Case (a). Using suitable trigonometric identities,

$$\sin(2x)\cos(3x) = \tfrac{1}{2}(\sin(2x+3x) + \sin(2x-3x))$$
$$= \tfrac{1}{2}\sin(5x) - \tfrac{1}{2}\sin x.$$

It follows that the Fourier coefficients are all zero except for $b_1 = -1/2$, $b_5 = 1/2$.

Case (b). Similarly, from
$$\cos^4 x = (\cos^2 x)^2 = \left(\tfrac{1+\cos(2x)}{2}\right)^2$$
$$= \tfrac{1}{4}\left(1 + 2\cos(2x) + \cos^2(2x)\right)$$
$$= \tfrac{1}{4}\left(1 + 2\cos(2x) + \tfrac{1+\cos(4x)}{2}\right)$$
$$= \tfrac{3}{8} + \tfrac{1}{2}\cos(2x) + \tfrac{1}{8}\cos(4x), \quad x \in \mathbb{R}.$$
Hence $a_0 = 3/4$, $a_2 = 1/2$, $a_4 = 1/8$, and the other coefficients are all zero. $\square$

7.B.5. Given the Fourier series of a function f on the interval $[-\pi, \pi]$ with coefficients a_m, b_n, $m \in \mathbb{N}$, $n \in \mathbb{Z}^+$, prove the following statements:

(a) If $f(x) = f(x+\pi)$, $x \in [-\pi, 0]$, then $a_{2k-1} = b_{2k-1} = 0$ for every $k \in \mathbb{N}$.

(b) If $f(x) = -f(x+\pi)$, $x \in [-\pi, 0]$, then $a_0 = a_{2k} = b_{2k} = 0$ for every $k \in \mathbb{N}$.

Solution. The case (a). For any $k \in \mathbb{N}$, the statement can be proved directly by calculations, but we provide here a conceptual explanation.

The definition of the function f ensures it is periodic with period π. Thus we may write its Fourier series on the shorter interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ as follows

$$f(t) = \frac{\tilde{a}_0}{2} + \sum_{n=1}^{\infty}\left(\tilde{a}_n \cos(2nt) + \tilde{b}_n \sin(2nt)\right).$$

Clearly this must be also the Fourier series of the same function f over the interval $[-\pi, \pi]$ and so the claim is proved.

Alternatively, if

$$f(x) = a_0/2 + \sum_{n=1}^{\infty}\left(a_n \cos(nx) + b_n \sin(nx)\right)$$

then

$$f(x+\pi) = a_0/2 + \sum_{n=1}^{\infty}\left(a_n \cos(nx+n\pi) + b_n \sin(nx+n\pi)\right)$$

$$= a_0/2 + (-1)^n \sum_{n=1}^{\infty}\left(a_n \cos(nx) + b_n \sin(nx)\right)$$

The two series are the same only when the odd coefficients are zero.

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 $m \neq n$,

$$\int_{-\pi}^{\pi} e^{imx}\, e^{-inx}\, dx = \int_{-\pi}^{\pi} e^{i(m-n)x}\, dx$$
$$= \tfrac{1}{i(m-n)}\left[e^{i(m-n)x}\right]_{-\pi}^{\pi} = 0.$$

Thus for $m \neq n$, the integral $\langle e^{imx}, e^{inx}\rangle = 0$.

FOURIER'S COMPLEX ORTHOGONAL SYSTEM

$$e^{-n\omega t}, \ldots, e^{-\omega t}, 1, e^{\omega t}, e^{2\omega t} \ldots, e^{n\omega t}, \ldots$$

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

$$\int_{-\pi}^{\pi}(\cos mx \cos nx + \sin mx \sin nx)\, dx = 0$$

$$\int_{-\pi}^{\pi}(\sin mx \cos nx - \cos mx \sin nx)\, dx = 0$$

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

$$\int_{-\pi}^{\pi} \cos^2 nx\, dx = \|\cos(nx)\|_2^2 = \pi,$$

$$\int_{-\pi}^{\pi} \sin^2 nx\, dx = \|\sin(nx)\|_2^2 = \pi,$$

If $n = 0$, then $\|1\|_2^2 = 2\pi$.

The case (b). Similarly if

$$f(x) = a_0/2 + \sum_{n=1}^{\infty} \big(a_n \cos(nx) + b_n \sin(nx) \big)$$

then $-f(x+\pi)$

$$= -a_0/2 - \sum_{n=1}^{\infty} \big(a_n \cos(nx + n\pi) + b_n \sin(nx - +n\pi) \big)$$

$$= -a_0/2 + (-1)^{n+1} \sum_{n=1}^{\infty} \big(a_n \cos(nx) + b_n \sin(nx) \big).$$

The two series are the same only when the even coefficients are zero. $\qquad\square$

Complex Fourier series. It is sometimes convenient (and often 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

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-inx} g(x) \, dx.$$

See the explanation in 7.1.7. We need just one formula for c_n, rather than one for a_n and another one for b_n.

7.B.6. Compute the complex version of the Fourier series $F(x)$ of the 2π–periodic function $g(x)$ defined by $g(x) = 0$, if $-\pi < x < 0$, while $g(x) = 1$ if $0 < x < \pi$.

Solution. We have for $n \neq 0$,

$$c_n = \frac{1}{2\pi} \int_0^{\pi} e^{-inx} \, dx = \frac{1}{2\pi} \left[\frac{e^{-inx}}{-in} \right]_0^{\pi} = \frac{1}{2\pi} \frac{(1-(-1)^n)}{in}.$$

while $c_0 = \frac{1}{2\pi} \int_0^{\pi} dx = 1/2$. So

$$F(x) = \frac{1}{2} + \sum_{n=1}^{\infty} c_n e^{inx} + \sum_{n=-\infty}^{-1} c_n e^{inx}$$

$$= \frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{2\pi} \frac{(1-(-1)^n)}{in} e^{inx} + \sum_{n=-\infty}^{-1} \frac{1}{2\pi} \frac{(1-(-1)^n)}{in} e^{inx}.$$

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

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-i\omega nt} \, dt.$$

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. $c_0 = a_0/2$, and for natural numbers n, we have

$$c_n = \tfrac{1}{2}(a_n - ib_n), \quad c_{-n} = \tfrac{1}{2}(a_n + ib_n).$$

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 2π, the Bessel inequality in this notation becomes

$$(1/2)|a_0|^2 + \sum_{n=1}^{\infty} (|a_n|^2 + |b_n|^2) \leq \|f\|_2^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt.$$

We have expressed the Fourier series $F(t)$ for a function $f(t)$ in the form

$$F(t) = \sum_{n=-\infty}^{\infty} c_n e^{i\omega nt}.$$

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.

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

$$g(x) = \frac{1}{2} \Big(\lim_{y \to x_+} f(y) + \lim_{y \to x_-} f(y) \Big).$$

(2) Moreover, if f is a continuous periodic function with a piecewise continuous derivative, then the pointwise convergence of its Fourier series is uniform.

(3) The L_2–distance $\|s_N - f\|_2$ converges to zero for $N \to \infty$.

In the second of the sums, replace n with $-n$. The second sum is then $\sum_{n=1}^{\infty} \frac{1}{2\pi} \frac{(1-(-1)^n)}{-in} e^{-inx}$ so that

$$F(x) = \frac{1}{2} + \frac{1}{2\pi} \sum_{n=1}^{\infty} \frac{(1-(-1)^n)}{in}(e^{inx} - e^{-inx})$$

$$= \frac{1}{2} + \frac{1}{2\pi} \sum_{n=1}^{\infty} \frac{(1-(-1)^n)}{in} 2i \sin(nx)$$

$$= \frac{1}{2} + \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{(1-(-1)^n)}{n} \sin(nx).$$

The terms when n is even are all 0. For n odd, put $n = 2m-1$ to obtain

$$F(x) = \frac{1}{2} + \frac{2}{\pi} \sum_{m=1}^{\infty} \left(\frac{1}{2m-1}\right) \sin(2m-1)x$$

$$= \frac{1}{2} + \frac{2}{\pi}\left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \dots\right).$$

Clearly, this is similar to the one problem of 7.1.9 which also could have been used directly to obtain the result just by an easy transformation. $\square$

7.B.7. Expand the following functions $g(x)$ on the interval $[-\pi, \pi]$ into the Fourier series, in the form $F(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$.

(i) $g(x) = x$, if $-\pi < x < \pi$.

(ii) $g(x) = |x|$, if $-\pi < x < \pi$.

(iii) $g(x) = x^2$, if $-\pi < x < \pi$.

(iv) $g(x) = 1$ if $-1/2 \le x \le 1/2$ and $g(x) = 0$ otherwise.

(v) $g(x) = 0$ if $-\pi < x < 0$, and $g(x) = x$ if $0 \le x \le \pi$.

$\bigcirc$

7.B.8. Repeat the task from the previous problem with the functions g defined again by the formulas (i) through (v), but taken on the interval $0 < x < 2\pi$. $\bigcirc$

7.B.9. Determine the convergence and uniform convergence of the Fourier series for the function $g(x) = e^{-x}$, $x \in [-1, 1)$.

Solution. It is not necessary to calculate the corresponding Fourier series if we wish only to check convergence (for the similar calculation in the case of function e^x see 7.B.13). Define the function $s(x)$ on $\mathbb{R}$ with period $T = 2$ as follows:

$$s(x) = g(x) = e^{-x}, \ x \in (-1, 1),$$

$$s(1) = \frac{g(-1) + \lim_{x \to 1-} g(x)}{2} = \frac{e + e^{-1}}{2}.$$

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.
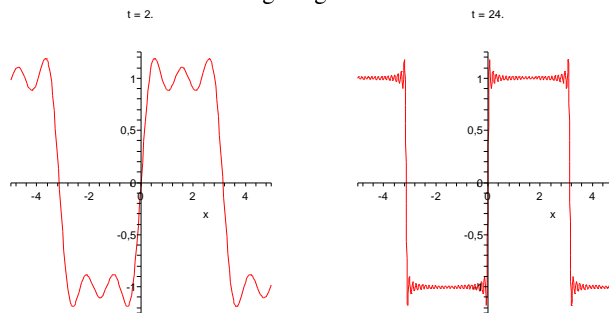
If the interval $[-T/2, T/2]$ is chosen for the prime period T of such a square wave function, the resulting Fourier series is

$$g(x) = \frac{4}{\pi}\left(\sin(\omega x) + \frac{1}{3}\sin(3\omega x) + \frac{1}{5}\sin(5\omega x) + \dots\right).$$

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.

This function is the sum of the Fourier series in question, cf. Theorem 7.1.8. In other words, the Fourier series converges to the periodic function $s(x)$. 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 $-1 < c < d < 1$. $\qquad \square$

7.B.10. Determine the cosine Fourier series (see the definitions at the end of the paragraph 7.1.10) for a periodic extension of the function

$$g(x) = 1, \ x \in [0, 1), \quad g(x) = 0, \ x \in [1, 4).$$

Determine also the sine Fourier series for

$$f(x) = x - 1, \ x \in (0, 2), \quad f(x) = 3 - x, \ x \in [2, 4).$$

Solution. We have already encountered the construction of a cosine Fourier series in 7.B.4(b) and also 7.B.3(b). It is the case of the Fourier series of an even function. Therefore, the first thing we to 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, $n \in \mathbb{N}$

$$a_0 = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \, dx = \frac{2}{8} \int_{-1}^{1} 1 \, dx = \frac{1}{2} \int_{0}^{1} 1 \, dx = \frac{1}{2},$$

$$a_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \cos(n\omega x) \, dx = \frac{1}{2} \int_{0}^{1} \cos \frac{n\pi x}{4} \, dx$$
$$= \frac{2}{n\pi} \sin \frac{n\pi}{4}.$$

We use

$$(1) \qquad \int_{-a}^{a} f(x) \, dx = 2 \int_{0}^{a} f(x) \, dx,$$

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 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). However, the converengence is uniform on any subinterval where the original function is continuous.

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:

$$a_n = \frac{2}{2} \int_{-1}^{1} x^2 \cos\left(\frac{2\pi n x}{2}\right) dx = 2 \int_{0}^{1} x^2 \cos(\pi n x) \, dx$$
$$= \frac{4}{\pi^2 n^2} (-1)^n.$$

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

$$g(x) = \frac{1}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos(\pi n x).$$

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, we obtain the remarkable result

$$\frac{\pi^2}{4} \left(g(1) - \frac{1}{3}\right) = \frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{(-1)^{2n}}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

We proceed with a further illustration. Because of the uniform convergence, we can differentiate term by term and calculate the Fourier series on the interval $-1 < x < 1$, for the function x

$$x = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(\pi n x).$$

This series cannot converge uniformly since the periodic extension of the function x from the interval $[-1, 1]$ is not a continuous function. However, it does converge pointwise

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 $n \in \mathbb{N}$

$$
\begin{aligned}
b_n &= \tfrac{2}{T} \int_{x_0}^{x_0+T} f(x) \sin (n\omega x) \, \mathrm{d}\, x \\
&= \tfrac{1}{2} \left(\int_0^2 (x-1) \sin \tfrac{n\pi x}{4} \, \mathrm{d}\, x - \int_2^4 (x-3) \sin \tfrac{n\pi x}{4} \, \mathrm{d}\, x \right) \\
&= \left[(1-x)\tfrac{2}{n\pi} \cos \tfrac{n\pi x}{4} \right]_0^2 + \left[\tfrac{8}{n^2\pi^2} \sin \tfrac{n\pi x}{4} \right]_0^2 \\
&\quad - \left[(3-x)\tfrac{2}{n\pi} \cos \tfrac{n\pi x}{4} \right]_2^4 - \left[\tfrac{8}{n^2\pi^2} \sin \tfrac{n\pi x}{4} \right]_2^4 \\
&= \tfrac{2}{n\pi} \left((-1)^n - 1 \right) + \tfrac{16}{n^2\pi^2} \sin \tfrac{n\pi}{2}.
\end{aligned}
$$

Immediately, $b_n = 0$ when n is even. So the sine Fourier series can be written

$$
\begin{aligned}
&\sum_{n=1}^{\infty} \left(\left(\tfrac{2}{n\pi} \left((-1)^n - 1 \right) + \tfrac{16}{n^2\pi^2} \sin \tfrac{n\pi}{2} \right) \sin \tfrac{n\pi x}{4} \right) \\
&= \sum_{n=1}^{\infty} \left(\left(\tfrac{-4}{(2n-1)\pi} + \tfrac{(-1)^{n-1}16}{(2n-1)^2\pi^2} \right) \sin \tfrac{(2n-1)\pi x}{4} \right). \quad \square
\end{aligned}
$$

7.B.11. Express the function $g(x) = \cos x$, $x \in (0, \pi)$, as the cosine Fourier series and the sine Fourier series.

Solution. Start with the sine series. This is the odd extension of the cosine function.

Necessarily, $a_n = 0$, $n \in \mathbb{N} \cup \{0\}$ for the sine series. $b_1 = \tfrac{2}{\pi} \int_0^{\pi} \cos x \sin x \, \mathrm{d}\, x = \tfrac{1}{\pi} \int_0^{\pi} \sin(2x) \, \mathrm{d}\, x = 0$. For all $n \geq 2$,

$$
\begin{aligned}
b_n &= \tfrac{2}{\pi} \int_0^{\pi} \cos x \sin(nx) \, \mathrm{d}\, x \\
&= \tfrac{1}{\pi} \int_0^{\pi} \sin((n+1)x) + \sin((n-1)x) \, \mathrm{d}\, x \\
&= -\tfrac{1}{\pi} \left[\tfrac{\cos((n+1)x)}{n+1} + \tfrac{\cos((n-1)x)}{n-1} \right]_0^{\pi} = \tfrac{2n\left((-1)^n + 1 \right)}{(n^2-1)\pi}.
\end{aligned}
$$

So $b_n = 0$ for odd $n \in \mathbb{N}$, and $b_n = \tfrac{4n}{(n^2-1)\pi}$ for even n. We conclude that

$$\cos x = \sum_{n=1}^{\infty} \left(\tfrac{8n}{(4n^2-1)\pi} \sin(2nx) \right), \quad x \in (0, \pi).$$

Of course, for the even extension of the function g in question,

$$g(x) = \cos x, \quad x \in (-\pi, \pi).$$

Thus the right hand side is the uniquely given cosine Fourier series. $\square$

$-1 < x < 1$, (see our reasonings about alternating series in 5.4.6), thus we have verified the above equality.

Similarly, we can integrate the Fourier series of x^2 term by term obtaining

$$\frac{1}{3}x^3 = \frac{2}{3}x + \frac{4}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^3} \sin(\pi n x).$$

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 intervat $[-1, 1]$ this way.

In this context, we use the following terminology:

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, illustrations, 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 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*[1] ψ, if possible with compact support, from which we create countably many functions $\psi_{j,k}$,

---

[1]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

7.B.12. Write the Fourier series of the π-periodic function which equals cosine on the interval $(-\pi/2, \pi/2)$. Then write the cosine Fourier series of the 2π-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

$$
\begin{aligned}
a_n &= \tfrac{2}{\pi} \int\limits_{-\pi/2}^{\pi/2} \cos x \cos(2nx) \, \mathrm{d}\,x \\
&= \tfrac{2}{\pi} \int\limits_{-\pi/2}^{\pi/2} \tfrac{1}{2}\big(\cos((2n+1)x) + \cos((2n-1)x)\big)\, \mathrm{d}x \\
&= \tfrac{1}{\pi} \left[\tfrac{\sin((2n+1)x)}{2n+1} + \tfrac{\sin((2n-1)x)}{2n-1} \right]_{-\pi/2}^{\pi/2} \\
&= \tfrac{2}{\pi} \left(\tfrac{(-1)^n}{2n+1} + \tfrac{(-1)^{n+1}}{2n-1} \right) \\
&= \tfrac{4}{\pi} \tfrac{(-1)^{n+1}}{4n^2-1}
\end{aligned}
$$

for every $n \in \mathbb{N}$. The calculation is also valid for $n = 0$, thus $a_0 = 4/\pi$. The desired Fourier series is

$$
\tfrac{2}{\pi} + \tfrac{4}{\pi} \sum_{n=1}^{\infty} \left(\tfrac{(-1)^{n+1}}{4n^2-1} \cos(2nx) \right).
$$

$\square$

7.B.13. Expand the function $g(x) = \mathrm{e}^x$ into

(a) a Fourier series on the interval $[0, 1)$;

(b) a cosine Fourier series on the interval $[0, 1]$;

(c) a sine Fourier series on the interval $(0, 1]$.

Solution. 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 $n = 10$.
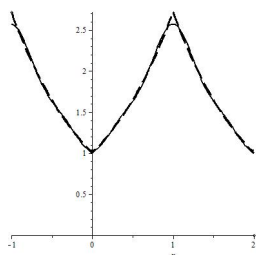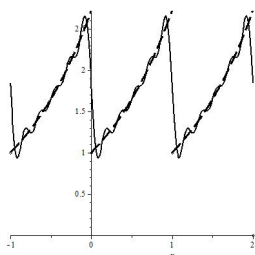
$j, k \in \mathbb{Z}$, by integer translations and dyadic dilations:

$$
\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).
$$

It is wise to rescale and choose ψ 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 ψ should match and cover all typical local behaviour of the data to be processed. We say that ψ is an *orthogonal mother wavelet* if the resulting countable system of functions ψ_{jk} 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 ψ 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.

Now we consider the question of effective treatment. Notice that we can also use the characteristic function φ of the interval $[0, 1)$ and write

$$
\psi(x) = \varphi(2x) - \varphi(2x-1) = \tfrac{1}{\sqrt{2}}\sqrt{2}\varphi_{1,0}(x) - \tfrac{1}{\sqrt{2}}\sqrt{2}\varphi_{1,1}(x).
$$

The function φ plays the role of the *father wavelet function* and it itself satisfies

$$
\varphi(x) = \varphi(2x) + \varphi(2x-1) = \tfrac{1}{\sqrt{2}}\sqrt{2}\varphi_{1,0}(x) + \tfrac{1}{\sqrt{2}}\sqrt{2}\varphi_{1,1}(x).
$$

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 ψ and φ, since we can find their values recurrently in all dyadic points x. Indeed,

$$
\varphi(2^{j-1}n) = \varphi(2^j n) + \varphi(2^j n - 1).
$$

combination of the high pass filters and low pass filters. The reader is advised to consult extremely rich literature, if interested in more details
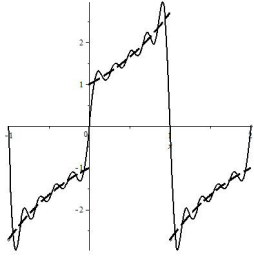
We use the formulae ($\alpha, \beta \in \mathbb{R}$)

(1) $\displaystyle \int \mathrm{e}^x \cos(\alpha x)\,\mathrm{d}x = \frac{\mathrm{e}^x\big(\alpha \sin(\alpha x) + \cos(\alpha x)\big)}{1+\alpha^2} + C,$

(2) $\displaystyle \int \mathrm{e}^x \sin(\beta x)\,\mathrm{d}x = \frac{\mathrm{e}^x\big(\sin(\beta x) - \beta \cos(\beta x)\big)}{1+\beta^2} + C,$

both of which can be obtained by two integrations by parts. Actually, the second one was computed in detail in 6.B.5(d). We obtain

(a) $\begin{aligned} a_n &= 2\int_0^1 \mathrm{e}^x \cos(2n\pi x)\,\mathrm{d}x \\ &= 2\left[\frac{\mathrm{e}^x\big(2n\pi \sin(2n\pi x) + \cos(2n\pi x)\big)}{1+4n^2\pi^2}\right]_0^1 \\ &= \frac{2(\mathrm{e}-1)}{1+4n^2\pi^2}, \quad n \in \mathbb{N}\cup\{0\}, \\ b_n &= 2\int_0^1 \mathrm{e}^x \sin(2n\pi x)\,\mathrm{d}x \\ &= 2\left[\frac{\mathrm{e}^x\big(\sin(2n\pi x) - 2n\pi \cos(2n\pi x)\big)}{1+4n^2\pi^2}\right]_0^1 \\ &= \frac{4n\pi(1-\mathrm{e})}{1+4n^2\pi^2}, \quad n \in \mathbb{N}; \end{aligned}$

(b) $\begin{aligned} a_n &= 2\int_0^1 \mathrm{e}^x \cos(n\pi x)\,\mathrm{d}x \\ &= 2\left[\frac{\mathrm{e}^x\big(n\pi \sin(n\pi x) + \cos(n\pi x)\big)}{1+n^2\pi^2}\right]_0^1 \\ &= \frac{2\big((-1)^n\mathrm{e}-1\big)}{1+n^2\pi^2}, \quad n \in \mathbb{N}\cup\{0\}; \end{aligned}$

(c) $\begin{aligned} b_n &= 2\int_0^1 \mathrm{e}^x \sin(n\pi x)\,\mathrm{d}x \\ &= 2\left[\frac{\mathrm{e}^x\big(\sin(n\pi x) - n\pi \cos(n\pi x)\big)}{1+n^2\pi^2}\right]_0^1 \\ &= \frac{2n\pi\big(1+(-1)^{n+1}\mathrm{e}\big)}{1+n^2\pi^2}, \quad n \in \mathbb{N}. \end{aligned}$

Substitution then yields the corresponding Fourier series for $g(x)$:

(a) $\mathrm{e}-1 + 2\,(\mathrm{e}-1)\sum_{n=1}^{\infty} \frac{\cos(2n\pi x)}{1+4n^2\pi^2}$

$\quad + 4\pi\,(1-\mathrm{e})\sum_{n=1}^{\infty} \frac{n\sin(2n\pi x)}{1+4n^2\pi^2};$

(b) $\mathrm{e}-1 + 2\sum_{n=1}^{\infty} \frac{\big((-1)^n\mathrm{e}-1\big)\cos(n\pi x)}{1+n^2\pi^2};$

(c) $2\pi \sum_{n=1}^{\infty} \frac{n\big(1+(-1)^{n+1}\mathrm{e}\big)\sin(n\pi x)}{1+n^2\pi^2}.$ □

The function φ 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 φ "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,

$$f(x) = \begin{cases} 0.3(x+3), & -2 \le x \le -1 \\ 0.7(x+1), & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

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 $|j| \le n$ and $|k| \le n$, the first one with $n = 5$, the second one with $n = 10$.
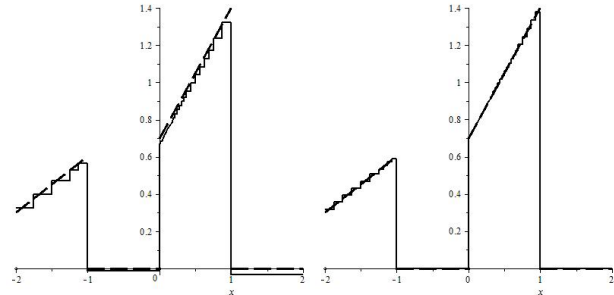
The approximation on the sides of the interval is not as good as in the middle, because we do not include enough shifts, i.e. values of k. One of the motivations for the scaling and shifting in the construction of wavelets is the hope for a small amount of non-zero coefficients. But this does not mean that most of the coefficients would be zero. In our case the percentage of non-zero coefficients for $n = \{1, 2, \ldots, 10\}$ 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 averadging and differencing relations based on certain natural requirements. Just as

464

7.B.14. 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

(1)
$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}, \qquad \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

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

$$f(x) = \tfrac{1}{3} + \tfrac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n \cos(n\pi x)}{n^2}, \quad x \in (-1, 1),$$

valid also for $x = \pm 1$. By adding π^2 and rescaling, it follows that

$$\begin{aligned} g(x) \quad &= \pi^2 - \left(\tfrac{1}{3} + \tfrac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n \cos\frac{n\pi x}{\pi}}{n^2} \right) \pi^2 \\ &= \tfrac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1}\cos(nx)}{n^2}, \quad x \in [-\pi, \pi]. \end{aligned}$$

Of course, one can also calculate the Fourier series of the function g directly.

Substituting $x = 0$ and $x = \pi$ then gives

$$\pi^2 = \tfrac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}, \quad \text{i.e.} \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12},$$

and

$$0 = \tfrac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1}(-1)^n}{n^2}, \quad \text{i.e.} \quad \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

In other words,

$$\begin{aligned} \pi^2 &= 12 \left(1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots \right) \\ &= 6 \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \right). \end{aligned}$$

$\square$

7.B.15. 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. For instance, the Fourier series

$$\frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2},$$

was calculated for the function $g(x) = |x|$, $x \in [-\pi, \pi)$, see 7.B.3(b). Since this function is continuous on $[-\pi, \pi)$ and $|-\pi| = |\pi|$,

$$|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos([2n-1]x)}{(2n-1)^2}, \quad x \in [-\pi, \pi].$$

Substituting $x = 0$ gives

$$0 = \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2}, \quad \text{hence} \quad \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8}.$$

$\square$

an indication, consider the slightly more general recurrent relations

$$\varphi(x) = \sqrt{2} \sum_{k=0}^{N} h_k \varphi(2x - k)$$

$$\psi(x) = \sqrt{2} \sum_{k=0}^{N} g_k \varphi(2x - k)$$

with yet unknown constants g_k and h_k. If we want the mother wavelet $\psi(x-k)$ to have zero coeffients 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 $k = 0, 1, \ldots, N-1$. 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.

Mother Wavelet

Father Wavelet
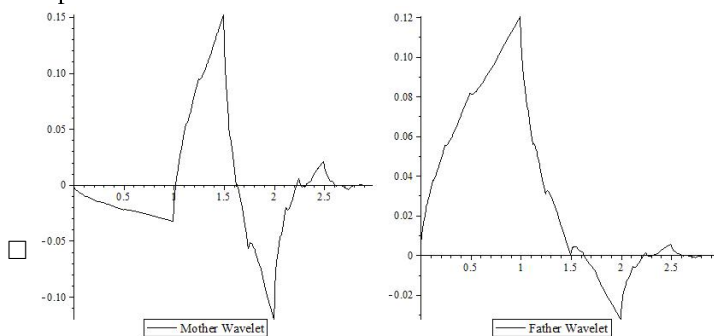
7.1.15. Proof of theorem 7.1.8 about Fourier series. We return to the detailed proof of the basic properties of the classical Fourier series. The reader could enjoy reading first the general context of metrics and convergence which we will introduce later in this chapter.

Thus, skipping the paragraphs up to 7.3.1, reading first there, 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.

7.B.16. Using the Fourier series of the function $g(x) = e^x$, $x \in [0, 2\pi)$, calculate $\sum_{n=1}^{\infty} \frac{1}{1+n^2}$.

Solution. (See (1), (2))

$$
\begin{aligned}
a_0 &= \frac{1}{\pi} \int\limits_0^{2\pi} e^x \, dx = \frac{1}{\pi} \left(e^{2\pi} - 1 \right), \\
a_n &= \frac{1}{\pi} \int\limits_0^{2\pi} e^x \cos(nx) \, dx = \frac{1}{\pi} \left[\frac{e^x [\cos(nx) + n \sin(nx)]}{1+n^2} \right]_0^{2\pi} \\
&= \frac{e^{2\pi} - 1}{(1+n^2)\pi}, \\
b_n &= \frac{1}{\pi} \int\limits_0^{2\pi} e^x \sin(nx) \, dx \\
&= \frac{1}{\pi} \left[\frac{e^x [\sin(nx) - n \cos(nx)]}{1+n^2} \right]_0^{2\pi} = -\frac{n(e^{2\pi} - 1)}{(1+n^2)\pi}.
\end{aligned}
$$

Therefore,

$$
e^x = \frac{e^{2\pi} - 1}{\pi} \left(\frac{1}{2} + \sum_{n=1}^{\infty} \frac{\cos(nx) - n \sin(nx)}{1+n^2} \right), \quad x \in (0, 2\pi).
$$

However, no choice of $x \in (0, 2\pi)$ yields the series $\sum_{n=1}^{\infty} \frac{1}{1+n^2}$ on the right-hand side. It would be obtained for $x = 0$. The periodic extension of g to $\mathbb{R}$ is not continuous at this point, so

$$
\begin{aligned}
\frac{e^0 + e^{2\pi}}{2} &= \frac{g(0) + \lim\limits_{x \to 2\pi^-} g(x)}{2} \\
&= \frac{e^{2\pi} - 1}{\pi} \left(\frac{1}{2} + \sum_{n=1}^{\infty} \frac{\cos 0 - n \sin 0}{1+n^2} \right),
\end{aligned}
$$

hence it follows that

$$
\frac{e^{2\pi} + 1}{2} \cdot \frac{\pi}{e^{2\pi} - 1} = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{1+n^2}
$$

which can be refined to

$$
\sum_{n=1}^{\infty} \frac{1}{1+n^2} = \frac{(\pi - 1)e^{2\pi} + \pi + 1}{2(e^{2\pi} - 1)}.
$$

$\square$

A few more cases of interesting series of numbers are shown in the exercises in the end of this Chapter, starting at the page 501. We add one more such exercise which reveals a different strategy.

7.B.17. 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.B.3(b). Parseval's identity

$$
\frac{a_0^2}{2} + \sum_{n=1}^{\infty} a_n^2 + \sum_{n=1}^{\infty} b_n^2 = \frac{2}{T} \int\limits_{x_0}^{x_0 + T} (g(x))^2 \, dx
$$

says, substituting for the a's and b's from our particular series, that

Although we need only the L_1 and L_2 norms now, we should observe that for general $1 \le p < \infty$, 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_∞ 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 the following bound on $\mathcal{S}^0[a, b]$. Namely, for $p > 1$ and $1/p + 1/q = 1$,

$$
\int_a^b |f(x)| \, dx \le |b - a|^{1/q} \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}
$$

$$
\le |b - a|^{1/q} \|f\|_p.
$$

Replace f with $f_n - f$. It is then clear from the above bound that L_p–convergence $f_n \to f$ 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 $\|f_n\| \le C$.

Then we can assert that $|f_n(x) - f(x)| \le 2C$, and it then follows that L_1–convergence implies L_p–convergence, since then

$$
\int_a^b |f(x) - f_n(x)|^p \, dx =
$$

$$
= \int_a^b |f(x) - f_n(x)|^{p-1} |f(x) - f_n(x)| \, dx
$$

$$
\le (2C)^{p-1} \int_a^b |f(x) - f_n(x)| \, dx
$$

which can be written

$$
\|f - f_n\|_p \le (2C)^{1/q} \|f - f_n\|_1^{1/p}.
$$

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.

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 refered to as the *Dirichlet condition*, which seems to have been derived as early as in 1824.

We begin by proving how this property of piecewise 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 $T = 2\pi$.

$\frac{\pi^2}{2} + \frac{16}{\pi^2} \sum\limits_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{1}{\pi} \int\limits_{-\pi}^{\pi} |x|^2 \, dx = \frac{2}{\pi} \int\limits_{0}^{\pi} x^2 \, dx = \frac{2\pi^2}{3},$

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 501. 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) $$y'' + \omega^2 y = 0$$

which arises in mechanics by Newton's law of force for a moving particle. Recall the brief introduction to the simplest differential equations in 6.2.14 on page 406. Much more follows in Chapter 8.

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 ω 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) $$z'' + \omega^2 z = 0.$$

The only difference between the equations (1) and (2) (besides the dissimilar physical interpretation) is the constant ω. 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) $$y'' + a^2 y = f(x)$$

with y an unknown in variable $x \in \mathbb{R}$, with a periodic, continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ on the right-hand

As the first step, we prepare simple bounds for the coefficients of the Fourier series. One bound is of course

$$|a_n| \le \frac{1}{\pi} \int_{-\pi}^{\pi} |f(x)| \, dx,$$

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 $\mathcal{S}^1[a, b]$, we can integrate by parts, thus obtaining

$$a_n(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) \, dx$$

$$= \frac{1}{n\pi} [f(x) \sin(nx)]_{-\pi}^{\pi} - \frac{1}{n\pi} \int_{-\pi}^{\pi} f'(x) \sin(nx) \, dx$$

$$= \frac{1}{n} b_n(f').$$

We write $a_n(f)$ for the corresponding coefficient of the function f, and so on.

Iterating this procedure, we really obtain a bound for functions f in $\mathcal{S}^{k+1}[-\pi, \pi]$ with continuous derivatives up to order k inclusive

$$|a_n(f)| \le \frac{1}{n^{k+1}\pi} \int_{-\pi}^{\pi} |f^{(k+1)}(x)| \, dx,$$

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 $f^{(k)}$.

Let f be a continuous function in the space $\mathcal{S}^1[a, b]$ such that the partial sums of its Fourier series converge pointwise to f. Then we can assert that

$$|s_N(x) - f(x)| = \left| \sum_{k=N+1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)) \right|$$

$$\le \sum_{k=N+1}^{\infty} (|a_k| + |b_k|).$$

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

$$|s_N(x) - f(x)| \le \sum_{k=N+1}^{\infty} \frac{1}{k} (|a_k'| + |b_k'|)$$

$$\le \left(\sum_{k=N+1}^{\infty} \frac{1}{k^2} \right)^{\frac{1}{2}} \left(\sum_{k=N+1}^{\infty} (|a_k'|^2 + 2|a_k'||b_k'| + |b_k'|^2) \right)^{\frac{1}{2}}$$

$$\le \left(\sum_{k=N+1}^{\infty} \frac{1}{k^2} \right)^{\frac{1}{2}} \left(\sum_{k=N+1}^{\infty} (2|a_k'|^2 + 2|b_k'|^2) \right)^{\frac{1}{2}}$$

$$\le \sqrt{2} \left(\int_{N}^{\infty} \frac{1}{x^2} \, dx \right)^{\frac{1}{2}} \frac{1}{\sqrt{\pi}} \|f'\|_2$$

467

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) $\quad f(x) = \dfrac{A_0}{2} + \sum_{n=1}^{\infty} \left(A_n \cos \dfrac{2\pi nx}{T} + B_n \sin \dfrac{2\pi nx}{T} \right).$

7.B.18. 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) $\qquad a \neq \dfrac{2\pi n}{T} \quad$ 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 $g, g', g'', \ldots$ 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)
$$g(x) = \dfrac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n \cos(\omega nx) + b_n \sin(\omega nx) \big), \quad x \in \mathbb{R},$$

where $\omega = 2\pi/T$. If g satisfies the equation (3), it has a continuous second derivative on $\mathbb{R}$. Therefore, for $x \in \mathbb{R}$,

$$g'(x) = \sum_{n=1}^{\infty} \big(\omega n b_n \cos(\omega nx) - \omega n a_n \sin(\omega nx) \big),$$

(3)
$$g''(x) = \sum_{n=1}^{\infty} \big(-\omega^2 n^2 a_n \cos(\omega nx) - \omega^2 n^2 b_n \sin(\omega nx) \big),$$

Substituting (4), (2) and (3) into (3) yields

$$a^2 \tfrac{1}{2} a_0 + \sum_{n=1}^{\infty} \big((-\omega^2 n^2 a_n + a^2 a_n) \cos(n\omega x) +$$
$$(-\omega^2 n^2 b_n + a^2 b_n) \sin(n\omega x) \big)$$
$$= \tfrac{A_0}{2} + \sum_{n=1}^{\infty} \big(A_n \cos(n\omega x) + B_n \sin(n\omega x) \big).$$

It follows that

(4) $\qquad a^2 \dfrac{a_0}{2} = \dfrac{A_0}{2}, \quad$ that is $\quad a^2 a_0 = A_0,$

$$= \left(\dfrac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2 \right) \cdot \dfrac{1}{\sqrt{N}}.$$

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:

$$\sup_{x \in \mathbb{R}} |s_N(x) - f(x)| \leq \left(\dfrac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2 \right) \cdot \dfrac{1}{\sqrt{N}}.$$

This proves the statement 7.1.8.(2), supposing the Dirichlet condition 7.1.8.(1) holds.

7.1.16. L_2–**convergence.** In the next step of our proof, we derive L_2–convergence of Fourier series under the condition of uniform convergence. 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.*

Here "dense" means that for any g in $\mathcal{S}^0[a,b]$ 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 δ satisfying $\pi > \delta > 0$, we define the function f_δ as x/δ for $|x| \leq \delta$ and $f_\delta(x) = h(x)$ otherwise. All the functions f_δ are continuous, in fact, piecewise linear. It can be calculated easily that $\|h - f_\delta\|_2 \to 0$ 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 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_ε 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 ε, and according to the assumption of uniform convergence for continuous functions, the second summand can be bounded also by ε. Notice that the third term has the value of the partial sum of the Fourier series for $f - f_\varepsilon$. Thus,

$$\|f - f_\varepsilon - s_N(f - f_\varepsilon)\| \leq \|f - f_\varepsilon\|.$$

and for $n \in \mathbb{N}$,

(5) $\quad \left(-\omega^2 n^2 + a^2 \right) a_n = A_n, \quad \left(-\omega^2 n^2 + a^2 \right) b_n = B_n.$

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

$$-\omega^2 n^2 + a^2 = -\left(\tfrac{2\pi n}{T} \right)^2 + a^2 \neq 0 \quad \text{for every } n \in \mathbb{N},$$

i.e., if (1) holds. In this case, the only solution of (3) with period T is determined by the only solution

(6) $\quad a_n = \dfrac{A_n}{-\omega^2 n^2 + a^2}, \quad b_n = \dfrac{B_n}{-\omega^2 n^2 + a^2}, \quad n \in \mathbb{N}$

of the system (5). We emphasize that we utilized the uniform convergence of the series in (3). $\qquad \square$

7.B.19. Using the solution of the previous problem, find all 2π-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.B.18, the condition $\sqrt{2} \notin \mathbb{N}$ implies that there is exactly one 2π-periodic solution. If we look for it as the value of the series

$$\tfrac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n \cos(nx) + b_n \sin(nx) \big), \quad x \in \mathbb{R},$$

we know also that (see (4) and (6))

$$a_0 = a_n = 0, \quad b_n = \tfrac{1}{n^2(2-n^2)}, \quad n \in \mathbb{N}.$$

Thus, the given equation has the unique 2π-periodic solution

$$y = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2(2-n^2)}, \quad x \in \mathbb{R}. \qquad \square$$

C. Convolution and Fourier Transform

Convolution is a typical integral operation used for smoothing data. See 7.2.2 for the definition and basic comments. It is defined by the formula

$$(f * g)(y) = \int_{-\infty}^{\infty} f(x) g(y - x) \, \mathrm{d}\, x.$$

The next problems introduces these features.

7.C.1. Find the convolutions $f * g$ and $f * h$ of the following functions and in each case check the "smoothing" of f.

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 continuous functions $s_N(f)$ to f which is we wanted to prove.

7.1.17. 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

$$s_N(t) = \frac{1}{T} \sum_{k=-N}^{N} \int_{-T/2}^{T/2} f(x) \, \mathrm{e}^{-i\omega k x} \, \mathrm{e}^{i\omega k t} \, dx,$$

where T is the period we are working with and $\omega = 2\pi/T$. This expression can be rewritten as

(1) $\qquad s_N(t) = \int_{-T/2}^{T/2} K_N(t - x) f(x) \, dx,$

where the function

$$K_N(y) = \frac{1}{T} \sum_{k=-N}^{N} \mathrm{e}^{i\omega k y}$$

is called the *Dirichlet kernel*. The sum is a (finite) geometric series with common ratio $\mathrm{e}^{i\omega y}$. By multiplying by $\mathrm{e}^{i\omega y}$ and then subtracting, we obtain

$$\mathrm{e}^{i\omega y} K_N(y) = \frac{1}{T} \sum_{k=-N}^{N} \mathrm{e}^{i\omega(k+1)y}$$

$$(1 - \mathrm{e}^{i\omega y}) K_N(y) = \frac{1}{T} \left(\mathrm{e}^{-iN\omega y} - \mathrm{e}^{i(N+1)\omega y} \right).$$

Provided ωy is not a multiple of 2π, we continue to obtain

$$\begin{aligned} K_N(y) &= \frac{1}{T} \frac{\mathrm{e}^{-iN\omega y} - \mathrm{e}^{i(N+1)\omega y}}{1 - \mathrm{e}^{i\omega y}} \\ &= \frac{1}{T} \frac{-\mathrm{e}^{-i(N+1/2)\omega y} + \mathrm{e}^{i(N+1/2)\omega y}}{\mathrm{e}^{i\omega y/2} - \mathrm{e}^{-i\omega y/2}} \\ &= \frac{1}{T} \frac{\sin((N+1/2)\omega y)}{\sin(\omega y/2)} \end{aligned}$$

in which the key step was to multiply both the numerator and the denominator by $-\mathrm{e}^{-i\omega y/2}$. 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

$$f(x) = \sin(x) + \frac{2}{5}\left[\sin(6x)\right]^2 - \frac{1}{5}\sin(60x)$$

$$g(x) = \begin{cases} \frac{1}{2\varepsilon} & -\varepsilon < x < \varepsilon \\ 0 & otherwise \end{cases}, \quad \varepsilon > 0$$

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \mu, \sigma \in \mathbb{R}, \sigma > 0.$$

Solution. The function g is chosen to provide the mean of f over (small) intervals of length 2ε 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:
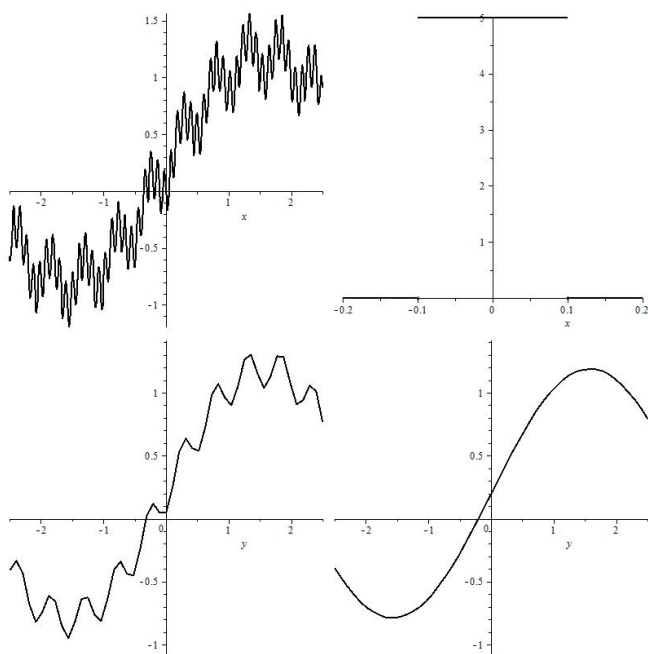
The graph depicted in the upper left is the original f, while the graph in the upper right is of g with $\varepsilon = 1/10$. The two lower graphs show their convolution with (respectively) parameters for g selected as $\varepsilon = 1/10$ and $\varepsilon = 13/50$. It is straightforward to compute the convolution explicitly, too:

$$f * g(y) = \frac{1}{2\varepsilon}\int_{y-\varepsilon}^{y+\varepsilon} \sin(x) + \frac{2}{5}\sin(6x)^2 - \frac{1}{5}\sin(60x)\,\mathrm{d}x$$

$$= \frac{1}{2\varepsilon}\left[-\cos(x) - \frac{1}{30}\sin(6x)\cos(6x) + \frac{1}{5}x + \frac{1}{300}\cos(60x)\right]_{y-\varepsilon}^{y+\varepsilon}.$$

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

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 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

$$s_N(x) - f(x) = \int_{-T/2}^{T/2} (f(x+y) - f(x))K_N(y)\,dy.$$

The integrand can be rewritten

$$\frac{f(x+y) - f(x)}{\sin(\omega y/2)}\sin((N+1/2)\omega y) =$$
$$= \varphi_x(y)(\cos(\omega y/2)\sin(N\omega y) + \sin(\omega y/2)\cos(N\omega y)),$$

where

$$\varphi_x(y) = \frac{f(x+y) - f(x)}{\sin(\omega y/2)}$$

for $y \neq 0$, while $\varphi_x(0) = 2f'(x)/\omega$. Since f is differentiable and continuous at the point x, the function $\varphi_x(y)$ is clearly continuous on the entire interval $[-T/2, T/2]$ (use L'Hospital's rule to see it).

Now, we can rewrite the integral expression for $s_N - f$ as

$$\frac{2}{T}\int_{-T/2}^{T/2} (\psi_1(y)\sin(N\omega y) + \psi_2(y)\cos(N\omega y)\,dy$$

with the continuous and bounded functions

$$\psi_1(y) = \frac{T}{2}\varphi_x(y)\cos(\omega y/2), \ \psi_2(y) = \frac{T}{2}\varphi_x(y)sin(\omega y/2).$$

Thus, we deal with the Fourier coefficients of the functions ψ_1 and ψ_2, which converge to zero. Hence

$$\lim_{N\to\infty}\int_{-T/2}^{T/2} \psi_1(y)\sin(N\omega y)\,dy = 0,$$

and

$$\lim_{N\to\infty}\int_{-T/2}^{T/2} \psi_2(y)\cos(N\omega y)\,dy = 0.$$

But this means $\lim s_N(x) - f(x) = 0$, as desired. See 7.1.5.(2)).

Now suppose the function f or its derivative has a discontinuity at $x = 0$. 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 $f(x) = f_1(x) + f_2(x)$,

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 $\mu = 0$, $\quad \sigma^2 = 1/60$ and (lower right) $\mu = 0$, $\sigma^2 = 5/60$.
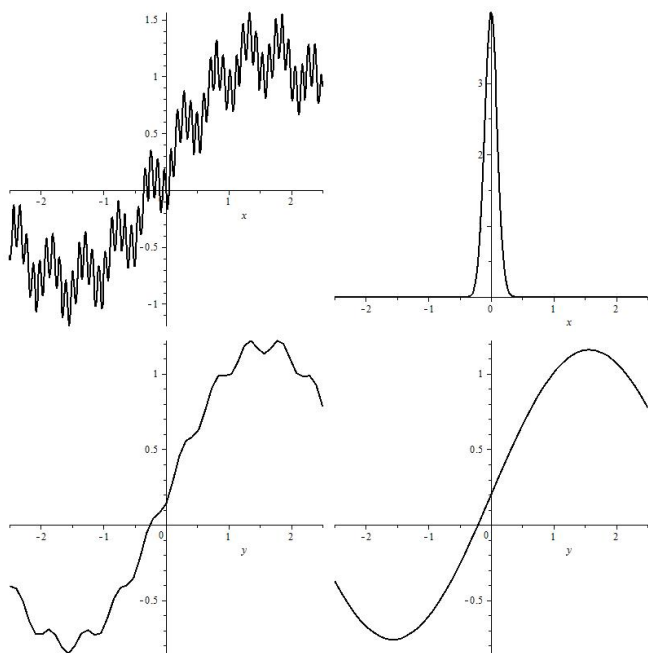
7.C.2. Determine the convolution $f_1 * f_2$ where

$$f_1(x) \;=\; \frac{1}{x} \qquad \text{for } x \neq 0$$
$$f_2(x) \;=\; \begin{cases} x & \text{for } x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Solution.

$$(f_1 * f_2)(t) = \int_{-\infty}^{\infty} f_1(x) f_2(t-x)\, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{x} f_2(t-x)\, \mathrm{d}x.$$

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

$$(f_1 * f_2)(t) = \int_{t-1}^{t+1} \frac{1}{x}(t - x)\, \mathrm{d}x = t \int_{t-1}^{t+1} \frac{1}{x}\, \mathrm{d}x - 2.$$

where

$$f_1(x) = \frac{1}{2}(f(x) + f(-x)) \quad \text{for} \quad x \neq 0$$
$$f_1(0) = \frac{1}{2}\big(\lim_{x \to 0_+} f(x) + \lim_{x \to 0_-} f(x) \big)$$
$$f_2(x) = \frac{1}{2}(f(x) - f(-x)).$$

Then the even part $f_1(x)$ is continuous and differentiable at the point $x = 0$ because of the existence of the one-sided limits, and so on a neighbourhood of the point $x = 0$. Also $f_2(0) = 0$, and the Fourier series for f_2 contains only the terms with $\sin(n\omega x)$.

Thus we can refer to the previous continuous case and obtain, for the Fourier series $F(x)$ of the function f, the equation

$$F(0) = \frac{1}{2}\big(\lim_{x \to 0_+} f(x) + \lim_{x \to 0_-} f(x) \big) + 0,$$

which is what we wanted to prove.

In the case of discontinuity at a point other that $x = 0$, we can proceed similarly. 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

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 α between vector spaces maps vectors to scalars. It is called a linear functional. 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:

$$(x_1, \ldots, x_n)^T \mapsto (\alpha_1, \ldots, \alpha_n) \cdot (x_1, \ldots, x_n)^T.$$

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 $I = [a, b]$. Linear mappings $\mathcal{S} \to \mathbb{R}$ are called (real) *linear functionals*. Examples of such functionals can be given in two different ways — 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.,

$$L(f) = f(x_0).$$

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 $\int_{t-1}^{t+1} \frac{1}{x} \, \mathrm{d}x$ with

$$\int_{\varepsilon}^{t+1} \frac{1}{x} \, \mathrm{d}x + \int_{t-1}^{-\varepsilon} \frac{1}{x} \, \mathrm{d}x$$

which computes to $\ln|t+1| - \ln|\varepsilon| + \ln|-\varepsilon| - \ln|t-1|$. The terms in ε cancel, so when we take the limit $\varepsilon \longrightarrow 0$, 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.C.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 $f_1(x) = 0$,

$$f_1 * f_2(t) = \int_{-\infty}^{\infty} f_1(x) f_2(t - x) \, \mathrm{d}x$$

$$= \int_{-1}^{1} (1 - x^2) f_2(t - x) \, \mathrm{d}x$$

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

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. 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 ε–neighbourhood of the origin. Similarly, we can work with the function

$$g(x) = \begin{cases} 0 & \text{if } |x| \geq \varepsilon \\ \mathrm{e}^{\frac{1}{x^2 - \varepsilon^2} + \frac{1}{\varepsilon^2}} & \text{if } |x| < \varepsilon \end{cases}$$

which we used in the paragraph 6.1.5. 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

$$g(x) = \frac{1}{\sqrt{\pi}} \, \mathrm{e}^{-x^2},$$

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.

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 $y \in \mathbb{R}$:

$$L_y(f) = \int_{-\infty}^{\infty} f(x) g(y - x) \, dx$$

472

$[-1, 1]$. Consequently,

$(f_1 * f_2)(t)$

$= 0, \text{ if } \quad t > 2$

$= \int_{t-1}^{1} (1 - x^2)(t - x)dx = 4t/3 - t^2 + t^4/12,$

$\qquad\qquad 1 \leq t \leq 2$

$= \int_{t-1}^{t} (1 - x^2)(t - x)dx = -t^2/2 + 1/4 + 2t/3,$

$\qquad\qquad 0 \leq t \leq 1$

$= \int_{-1}^{t} (1 - x^2)(t - x)dx$

$= -t^4/12 + t^2/2 + 1/4 + 2t/3,$

$\qquad\qquad -1 \leq t \leq 0$

$= 0, \quad \text{ if } \quad t < -1.$

$\square$

7.C.4. Determine the convolution $f_1 * f_2$ of the functions

$$f_1 = \begin{cases} 1 - x & \text{for } x \in [-2, 1], \\ 0 & \text{otherwise}, \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{for } x \in [0, 1], \\ 0 & \text{otherwise}. \end{cases}$$

$\bigcirc$

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.C.5. Fix $\Omega > 0$. Recall that $\operatorname{sgn} t = 1$ if $t > 0$, and $\operatorname{sgn} t = -1$ if $t < 0$, $\operatorname{sgn} 0 = 0$.

Find the Fourier transform $\mathcal{F}(f)$ and the inverse Fourier transform $\mathcal{F}^{-1}$ of the functions:

(a) $f(t) = \operatorname{sgn} t$ 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

$$\mathcal{F}(f)(\omega) = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t}\, \mathrm{d}t$$

$$= \tfrac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \operatorname{sgn} t \,\big(\cos(\omega t) - i\sin(\omega t)\big)\, \mathrm{d}t$$

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.

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

$$* : \mathcal{S}_c \times \mathcal{S}_c \to \mathcal{S}_c$$

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 modelling 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 useful case of convolution. In paragraph 7.1.17, 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 $N = 15$.
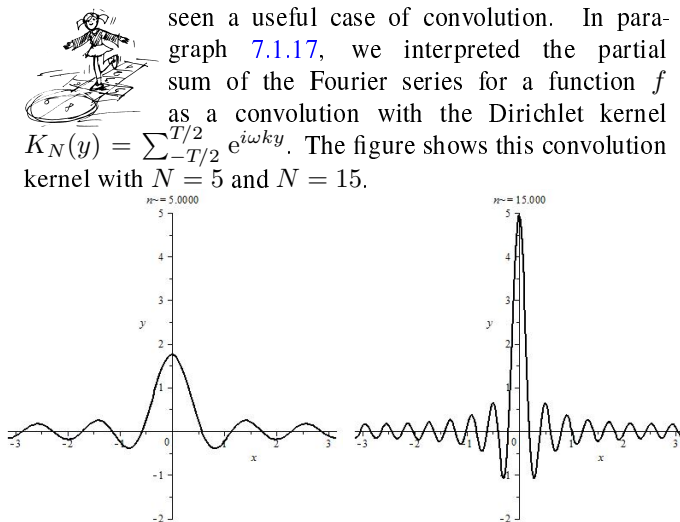
$$= \tfrac{1}{\sqrt{2\pi}} \left(\int_0^{\Omega} \big(\cos(\omega t) - i\sin(\omega t)\big) \, \mathrm{d}\,t - \int_{-\Omega}^{0} \big(\cos(\omega t) - i\sin(\omega t)\big) \, \mathrm{d}\,t \right).$$

Since cos and sin are respectively even and odd functions,

$$\mathcal{F}(f)(\omega) = \tfrac{2}{\sqrt{2\pi}} \int_0^{\Omega} -i\sin(\omega t) \, \mathrm{d}\,t = \tfrac{2i}{\sqrt{2\pi}} \left[\tfrac{\cos(\omega t)}{\omega} \right]_0^{\Omega}$$

$$= i\sqrt{\tfrac{2}{\pi}} \, \tfrac{\cos(\omega\Omega) - 1}{\omega}.$$

The inverse Fourier transforms 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 ω. Thus, the only difference in the result is the sign:

$$\mathcal{F}^{-1}(f)(t) = \tfrac{2}{\sqrt{2\pi}} \int_0^{\Omega} i\sin(\omega t) \, \mathrm{d}\,\omega = \tfrac{-2i}{\sqrt{2\pi}} \left[\tfrac{\cos(\omega t)}{t} \right]_0^{\Omega}$$

$$= i\sqrt{\tfrac{2}{\pi}} \, \tfrac{1 - \cos(t\Omega)}{t}.$$

Case (b) is computed similarly:

$$\mathcal{F}(f)(\omega) = \tfrac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \big(\cos(\omega t) - i\sin(\omega t)\big) \, \mathrm{d}\,t$$

$$= \tfrac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \cos(\omega t) \, \mathrm{d}\,t = \tfrac{1}{\sqrt{2\pi}} \left[\tfrac{\sin(\omega t)}{\omega} \right]_{-\Omega}^{\Omega} = \sqrt{\tfrac{2}{\pi}} \, \tfrac{\sin(\omega\Omega)}{\omega}.$$

The latter expression is often expressed by means of the function $\operatorname{sinc}(t) = \sin(t)/t$

$$\mathcal{F}(f)(\omega) = \frac{2\Omega}{\sqrt{2\pi}} \operatorname{sinc}(\omega\Omega).$$

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:

$$\mathcal{F}^{-1}(f)(t) = \frac{2\Omega}{\sqrt{2\pi}} \operatorname{sinc}(t\Omega).$$

$\square$

The results are:
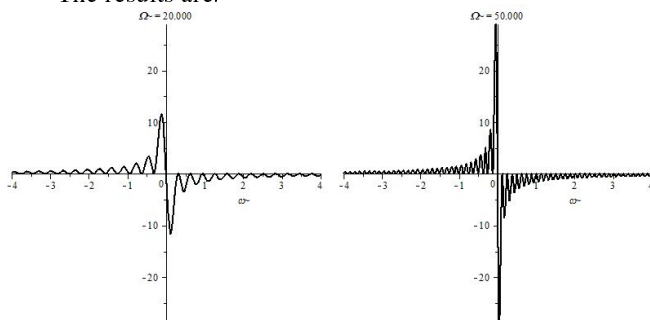
Notice that instead of integrals over the entire real line, we employ the integration over the basic period T of the periodic functions in question.

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 Gibs effect in general.

We do not provide more details here. Readers may either work them out themselves (as a nontrivial exercise) or look them up in the literature.

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,

$$k : \mathbb{R}^2 \to \mathbb{R},$$

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

$$f(t) = \sum_{n=-\infty}^{\infty} c_n \, e^{i\omega_n t},$$

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

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) \, e^{-i\omega_n t} \, dt$$
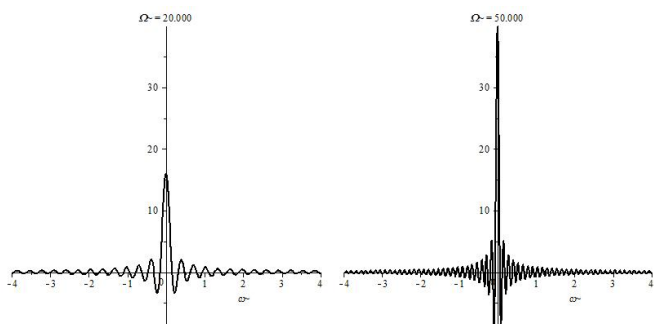
The first two diagrams below show the imaginary values of the Fourier image of the signum function from 7.C.5(a) with $\Omega = 20$ and $\Omega = 50$. The next two diagrams do the same for the characteristic function of the interval $|x| < \Omega$ from 7.C.5(b). The longer the interval with the constant 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.C.6. Discover how the Fourier transform and its inverse behave under the translation τ_a in the variable, $\tau_a f(x) = f(x + a)$, and the phase shift φ_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:

$$\mathcal{F} \circ \tau_a f(\omega) = \int_{-\infty}^\infty f(t + a) \, e^{-i\omega t} \, dt$$

$$= \int_{-\infty}^\infty f(x) \, e^{-i\omega(x - a)} \, dx = e^{ia\omega} \, \mathcal{F}f(\omega).$$

$$\mathcal{F} \circ \varphi_a f(\omega) = \int_{-\infty}^\infty f(t) \, e^{iat} \, e^{-i\omega t} \, dt$$

$$= \int_{-\infty}^\infty f(t) \, e^{-i(\omega - a)t} \, dt = \mathcal{F}f(\omega - a).$$

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 ω_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 $f(t)$ describes the summands of the Riemann sums for the improper integral

$$\frac{1}{2\pi} \int_{-\infty}^\infty g(\omega) \, e^{i\omega t} \, d\omega,$$

where $g(\omega)$ is a function which takes, at the points ω_n, the values

$$g(\omega_n) = \int_{-T/2}^{T/2} f(x) \, e^{-i\omega_n x} \, dx.$$

We are working with piecewise continuous functions with a compact support, thus our function f is integrable in absolute value over $\mathbb{R}$. Letting $T \to \infty$, 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 show that there is a large set of Riemann integrable functions f on $\mathbb{R}$ for which we can define a pair of mutually inverse integral operators:

FOURIER TRANSFORM

For every piecewise continuous real or complex function f on $\mathbb{R}$ with compact support, we define

$$\mathcal{F}(f)(\omega) = \tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty f(t) \, e^{-i\omega t} \, dt.$$

This function $\tilde{f}$ is called the Fourier transform of the function f. The previous ideas show that

$$f(t) = \mathcal{F}^{-1}(\tilde{f})(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \tilde{f}(\omega) \, e^{i\omega t} \, d\omega.$$

This says that the *Fourier transform* $\mathcal{F}$ just defined has an inverse operation $\mathcal{F}^{-1}$, which is called the *inverse Fourier transform*.

Notice that both the Fourier transform and its inverse are integral operators with almost identical kernels

$$k(\omega, t) = e^{\pm i\omega t}.$$

Of course, these transforms are meaningful for a much larger class of functions. Interested readers are referenced to specialized literature.

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,

$$\mathcal{F}^{-1} \circ \tau_a = \varphi_{-a} \circ \mathcal{F}^{-1}, \quad \mathcal{F}^{-1} \circ \varphi_a = \tau_a \circ \mathcal{F}^{-1}.$$

$\square$

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.C.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

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} e^{-i\omega t} \, dt.$$

A standard trick is to transform the problem into one of solving a (simple) differential equation. Hence:

Differentiating (with respect to ω) and then integrating by parts gives

$$\left(\mathcal{F}(f)(\omega) \right)' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -it \, e^{-at^2} e^{-i\omega t} \, dt$$

$$= \frac{1}{\sqrt{2\pi}} \left(\lim_{t \to \infty} \frac{i}{2a} e^{-at^2 - i\omega t} - \lim_{t \to -\infty} \frac{i}{2a} e^{-at^2 - i\omega t} \right.$$

$$\left. - \int_{-\infty}^{\infty} \frac{i(-i\omega)}{2a} e^{-at^2} e^{-i\omega t} \, dt \right)$$

$$= \frac{1}{\sqrt{2\pi}} \left(\frac{i}{2a} \lim_{t \to \infty} e^{-at^2} - \frac{i}{2a} \lim_{t \to -\infty} e^{-at^2} \right.$$

$$\left. - \int_{-\infty}^{\infty} \frac{\omega}{2a} e^{-at^2} e^{-i\omega t} \, dt \right)$$

$$= -\frac{\omega}{2a} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} e^{-i\omega t} \, dt \right) = -\frac{\omega}{2a} \mathcal{F}(f)(\omega).$$

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

$$\ln|y| = -\frac{\omega^2}{4a} - C, \quad \text{i.e.} \quad y = K e^{-\frac{\omega^2}{4a}},$$

where C and K are constants. All solutions (including the zero solution) of the differential equation are given by the function

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

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\omega t} \, d\omega = \frac{1}{\sqrt{2\pi}} \left[\frac{1}{it} e^{i\omega t} \right]_{-\Omega}^{\Omega}$$

$$= \frac{2}{\sqrt{2\pi} t} \frac{1}{2i} (e^{i\Omega t} - e^{-i\Omega t})$$

$$= \frac{2\Omega}{\sqrt{2\pi}} \frac{\sin(\Omega t)}{\Omega t}.$$

Thus, except for a multiplicative constant and the scaling of the input variable, it is the very important function $f(x) = \text{sinc}(x) = \frac{\sin x}{x}$.

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 $t = \pm\pi/\Omega$ 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 Ω increases is also depicted by a curve.

Omega = 20.000
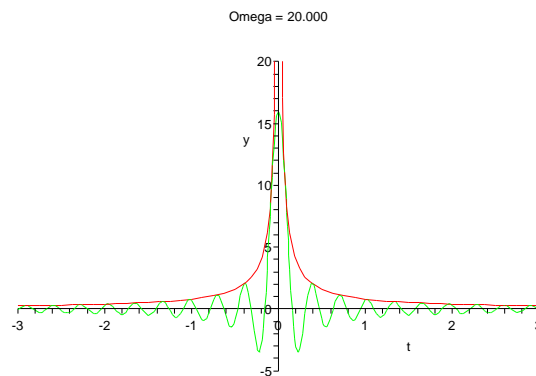
The indicator function of the interval $[-\Omega, \Omega]$ is Fourier-transformed to the function f, which has takes significant positive values near zero, and the value taken at zero is a fixed multiple of Ω. Therefore, as Ω increases, the f concentrates more and more near the origin.

Now we derive the Fourier transform of the derivative $f'(t)$ for a function f. We continue to suppose that f has compact support, so that both $\mathcal{F}(f')$ and $\mathcal{F}(f)$ exist. By integration by parts,

$$\mathcal{F}(f')(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} f'(t) e^{-i\omega t} \, dt$$

$$= \frac{1}{\sqrt{2\pi}} \left[e^{-i\omega t} f(t) \right]_{-\infty}^{\infty} + \frac{i\omega}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt$$

$$= i\omega \mathcal{F}(f)(\omega).$$

Thus the Fourier transform converts the (limit) operation of differentiation to the (algebraic) operation of a multiplication

$$y(\omega) = K\,\mathrm{e}^{-\frac{\omega^2}{4a}}, \quad K \in \mathbb{R}.$$

To find K, begin with the well known fact (proved in **??**)

$$\int\limits_{-\infty}^{\infty} \mathrm{e}^{-x^2}\,\mathrm{d}\,x = \sqrt{\pi},$$

to obtain

$$\int\limits_{-\infty}^{\infty} \mathrm{e}^{-at^2}\,\mathrm{d}\,t = \frac{1}{\sqrt{a}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-x^2}\,\mathrm{d}\,x = \frac{\sqrt{\pi}}{\sqrt{a}}.$$

Therefore, $\mathcal{F}(f)(0) = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\pi}}{\sqrt{a}} = \frac{1}{\sqrt{2a}}$ and $\mathcal{F}(f)(0) = K\,\mathrm{e}^0 = K$. So $K = \frac{1}{\sqrt{2a}}$ and

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2a}}\,\mathrm{e}^{-\frac{\omega^2}{4a}}.$$

$\square$

7.C.8. Determine the Fourier transform image of the Gaussian function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Solution. Use the result of the previous problem with $a = \frac{1}{2\sigma^2}$ and the composition with the variable shift τ_a from the last but one problem. It follows that

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{i\mu\omega}\,\mathrm{e}^{-\sigma^2\frac{\omega^2}{2}}.$$

$\square$

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_Ω of the interval $(-\Omega, \Omega)$.

7.C.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

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\Omega}^{\Omega}\cos(nt)\,\mathrm{e}^{-i\omega t}\,\mathrm{d}\,t$$

$$= \frac{1}{\sqrt{2\pi}}\int\limits_{-\Omega}^{\Omega}\frac{1}{2}(\mathrm{e}^{int}+\mathrm{e}^{-int})\,\mathrm{e}^{-i\omega t}\,\mathrm{d}\,t$$

$$= \frac{1}{2\sqrt{2\pi}}\Big[\frac{1}{i(n-\omega)}\,\mathrm{e}^{i(n-\omega)t}\Big]_{-\Omega}^{\Omega}$$

$$\qquad + \frac{1}{2\sqrt{2\pi}}\Big[\frac{-1}{i(n+\omega)}\,\mathrm{e}^{-i(\omega+n)t}\Big]_{-\Omega}^{\Omega}$$

$$= \frac{\Omega}{\sqrt{2\pi}}\big(\mathrm{sinc}((n-\omega)\Omega)+\mathrm{sinc}((n+\omega)\Omega)\big).$$

by the variable. Of course, this procedure can be iterated, to obtain

$$\mathcal{F}(f'')(\omega) = -\omega^2\mathcal{F}(f), \ldots, \mathcal{F}(f^{(n)}) = i^n\omega^n\mathcal{F}(f).$$

7.2.7. The relation to convolutions. There is another extremely important property to consider, namely the relation between convolutions and Fourier transforms. Calculate the transform of the convolution $h = f * g$, where, as usual, the functions are assumed to have compact support. Recall that we may change the order of integration, see 6.3.13. Then we change variable by the substitution $t - x = u$. The result is

$$\mathcal{F}(h)(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}f(x)g(t-x)\,dx\right)\mathrm{e}^{-i\omega t}\,dt$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}f(x)\left(\int_{-\infty}^{\infty}g(t-x)\,\mathrm{e}^{-i\omega t}\,dt\right)dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}f(x)\left(\int_{-\infty}^{\infty}g(u)\,\mathrm{e}^{-i\omega(u+x)}\,du\right)dx$$

$$= \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty}f(x)\,\mathrm{e}^{-i\omega x}\,dx\right)\cdot\left(\int_{-\infty}^{\infty}g(u)\,\mathrm{e}^{-i\omega u}\,du\right)$$

$$= \sqrt{2\pi}\mathcal{F}(f)\cdot\mathcal{F}(g)$$

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 $\mathrm{sinc}(t)$ 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. As an illustration of the power of our simple results, look at the behaviour of the Fourier transform with respect to the L_2–norm. We write $\hat{g}$ for the function $\hat{g}(t) = \overline{g(-t)}$ and notice that

$$(f*\hat{g})(t) = \int_{-\infty}^{\infty}f(x)\hat{g}(t-x)\,dx = \int_{-\infty}^{\infty}f(x)\overline{g(x-t)}\,dx.$$

In particular, the scalar product is given by the formula

$$\langle f, g\rangle = (f*\hat{g})(0).$$

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:

$$\mathcal{F}(g)(\omega) = i\frac{\Omega}{\sqrt{2\pi}}\big(-\operatorname{sinc}((n-\omega)\Omega) + \operatorname{sinc}((n+\omega)\Omega)\big)$$

$\square$

7.C.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 $\Omega \to \infty$?

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
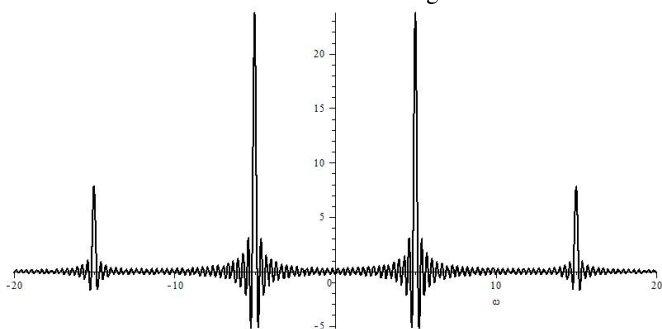
Each of the peaks behaves like the Fourier image of the characteristic function h_Ω, shifted to the frequencies.

If Ω increases to infinity, the image $\tilde{f}$ has four peaks at the same positions corresponding to the frequencies ± 5 and ± 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.C.11. Find the Fourier transform image of the convolution of the signals $f(t)$ and $g(t)$ from the Problem 7.C.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.C.5 and in the last two problems above, we already computed the Fourier

Now, the definition of the Fourier transform yields

$$\langle f, g \rangle = (\mathcal{F}^{-1}\mathcal{F}(f * \hat{g}))(0)$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathcal{F}(f * \hat{g})\,e^{ix\omega}\,d\omega_{|x=0}$$
$$= \int_{-\infty}^{\infty}\tilde{f}(\omega)\tilde{\hat{g}}(\omega)\,d\omega$$

while

$$\tilde{\hat{g}} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\overline{g(-x)}\,e^{-i\omega x}\,dx = \bar{\tilde{g}}.$$

Consequently, $\langle f, g \rangle = \langle \tilde{f}, \tilde{g} \rangle$. 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_Ω of the interval $[-\Omega, \Omega]$. Let Ω 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 Ω increases to ∞, 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* δ 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 δ 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. 6.3.19, 6.3.15 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 the Dirac δ. For instance,

$$\mathcal{F}(\cos(nt))(\omega) = \sqrt{\frac{\pi}{2}}\big(\delta(n-\omega) + \delta(n+\omega)\big),$$

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 $\Omega = 5, \varepsilon = 1/10$.
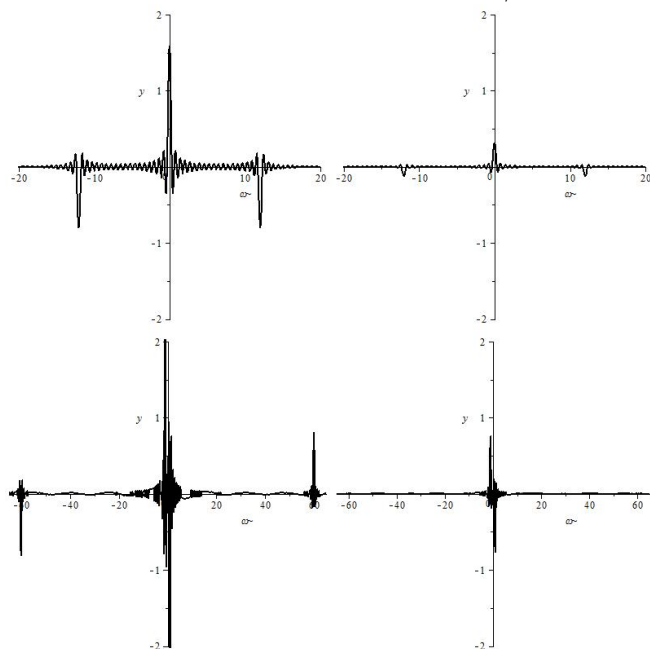
The reader should compare the diagrams of f and $f * g$ in 7.C.1 to see that the higher frequencies in f are effectively canceled by this convolution, as expected. □

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.C.12. By using the inverse Fourier transform solve the integral equation

$$\int_0^\infty f(t) \sin(xt) \, \mathrm{d}t = \mathrm{e}^{-x}, \quad x > 0$$

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

$$f(t) = \frac{2}{\pi} \int_0^\infty \mathrm{e}^{-x} \sin(xt) \, dx, \quad t > 0.$$

which can be seen from the calculation of the Fourier transform of the function $f_\Omega \cos(nx)$ and then letting Ω approach ∞, see the solution to problem 7.C.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.C.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.C.7, the Fourier image of the Gaussian function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-\frac{t^2}{2\sigma^2}}$$

is again Gaussian corresponding to the reciprocal values of σ. In the limit $\sigma \to 0$, the image converges fast to the multiple of constant function, see the illustrations. with $\sigma = 3$ and $\sigma = 1/10$.
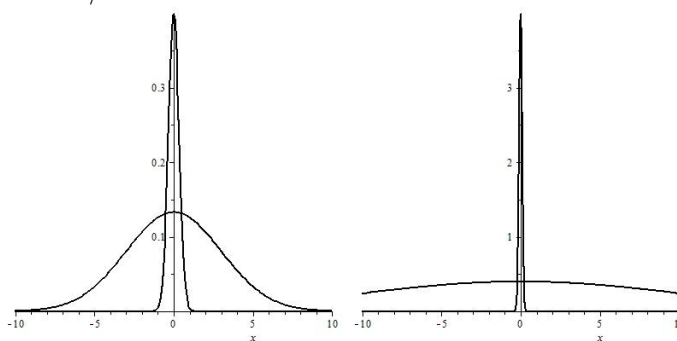
Notice that the rather large σ in the first illustration corresponds to a wide Gaussian, while the image is the slim one. The second illustration provides the opposite case. The preimage is the narrow Gaussian and the image is already reasonably close to the constant function. The Gaussians are chosen 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 $f(-t) = -f(t)$, the contribution in the integration of the product of

479

Integrating by parts twice shows that

$$\int e^{-x} \sin(xt) \, dx = \tfrac{e^{-x}}{1+t^2} \left[-\sin(xt) - t\cos(xt) \right] + C.$$

Hence

$$\int_0^\infty e^{-x} \sin(xt) \, dx$$

$$= \lim_{x \to \infty} \left(\tfrac{e^{-x}}{1+t^2} \left[-\sin(xt) - t\cos(xt) \right] \right) - \tfrac{e^0(-t)}{1+t^2} = \tfrac{t}{1+t^2}.$$

So

$$f(t) = \frac{2}{\pi} \frac{t}{1+t^2}, \quad t > 0.$$

$\square$

7.C.13. Use the Fourier transform to find solutions of the non-homogeneous linear differential equation

(1) $$y' = a\,y + f$$

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

$$\tilde{y} = \frac{1}{i\omega - a}\tilde{f}$$

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

$$\int_{-\infty}^0 e^{at}\,e^{-i\omega t}\,d\,t = \left[\tfrac{1}{a-i\omega}\,e^{(a-i\omega)t} \right]_{-\infty}^0 = \frac{1}{a - i\omega}.$$

Similarly for $a > 0$

$$\int_0^\infty e^{at}\,e^{-i\omega t}\,d\,t = \left[\tfrac{1}{a-i\omega}\,e^{(a-i\omega)t} \right]_0^\infty = \frac{1}{i\omega - a}.$$

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

$f(t)$ 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 determined similarly:

$$\tilde{\mathcal{F}}(f)(\omega) = \frac{2i}{\sqrt{2\pi}} \int_0^\infty f(t)\sin\omega t \, dt.$$

Omitting the imaginary unit i 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:

$$\tilde{f}_c(\omega) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(t)\cos(\omega t) \, dt,$$

$$f(t) = \sqrt{\frac{2}{\pi}} \int_0^\infty \tilde{f}_s(t)\sin\omega t \, dt.$$

7.2.11. Laplace transforms. The Fourier transform cannot be applied to functions which are not integrable in absolute value over $\mathbb{R}$ (at least, we do not obtain true functions). The *Laplace transform* is similar to the Fourier transform:

$$\mathcal{L}(f)(s) = \bar{f}(s) = \int_0^\infty f(t)\,e^{-st} \, dt.$$

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:

$$\mathcal{L}(f'(t))(s) = \int_0^\infty f'(t)\,e^{-st} \, dt$$

$$= [f(t)\,e^{-st}]_0^\infty + s \int_0^\infty f(t)\,e^{-st} \, dt$$

$$= -f(0) + s\mathcal{L}(f)(s).$$

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.D.1.

provides the requested Fourier image. Immediately it is seen that the convolution

$$y(t) = -\int_{-\infty}^{0} e^{ax} f(t-x) \, d x$$

is a solution. (The multiples $\sqrt{2\pi}$ in the expression with the convolution cancel.) Similarly, if $a < 0$ then

$$y(t) = \int_{0}^{\infty} e^{ax} f(t-x) \, d x$$

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

$$y(t) = -\int_{-\infty}^{0} e^{x}(t-x) \, d x = -t - 1.$$

$\square$

7.C.14. Check directly that the two functions $y(t)$ found above are indeed solutions to the equation $y' = a y + f$. $\bigcirc$

7.C.15. As in the previous problem, solve the second order equation

$$y'' = a y + f.$$

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

$$\tilde{y} = \frac{-1}{\omega^2 + a} \tilde{f}.$$

In order to guess the correct preimage of the rational function in question, first assume $a > 0$ and compute

$$\int_{-\infty}^{\infty} e^{-a|x|} e^{-i\omega x} \, d x$$

$$= \left[\tfrac{1}{a-i\omega} e^{a-i\omega x}\right]_{-\infty}^{0} + \left[\tfrac{-1}{a+i\omega} e^{-a-i\omega x}\right]_{0}^{\infty}$$

$$= \frac{1}{a-i\omega} + \frac{1}{a+i\omega} = \frac{2a}{a^2+\omega^2}$$

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)

$$y(t) = \frac{-1}{\sqrt{a}} e^{-\sqrt{a}|t|} * f(t) = \frac{-1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-\sqrt{a}|x|} f(t-x) \, d x.$$

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* τ 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 $r = 0, 1, \ldots, N - 1$. 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 should not be far from

$$\hat{\tilde{f}}(k) = \sum_{r=0}^{N-1} \tilde{f}(r) e^{i\frac{2\pi}{N}kr}.$$

Actually, these are already the mutually inverse transformations:

Theorem. *The transformation above satisfies $\hat{\tilde{f}}(k) = f(k)$ for all $k = 0, 1, \ldots, N - 1$.*

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 $\{0, 1, 2 \ldots N - 1\}$, $k - s$ 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:

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

$$y(t) = \frac{1}{\sqrt{-4a}} \int_{-\infty}^{\infty} \sin(\sqrt{-a}|x|) f(t - x) \, \mathrm{d}\, x.$$

If we rewrite the equation as $y'' + by = f$ with $b > 0$, the result says

$$y(t) = \frac{1}{2b} \int_{-\infty}^{\infty} \sin(b|x|) f(t - x) \, \mathrm{d}\, x.$$

$\square$

7.C.16. Check directly that the two functions $y(t)$ found above are indeed solutions to the equation $y'' = a\, y + f$. $\bigcirc$

D. Laplace Transform

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.D.1. 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) $f(t) = t^k$, $k \in \mathbb{N}$,

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).

$$\hat{\hat{f}}(k) = \sum_{r=0}^{N-1} \left(\frac{1}{N} \sum_{s=0}^{N-1} f(s) \, e^{-i \frac{2\pi}{N} rs} \right) e^{i \frac{2\pi}{N} rk}$$

$$= \frac{1}{N} \sum_{s=0}^{N-1} f(s) \left(\sum_{r=0}^{N-1} e^{i \frac{2\pi}{N} r(k-s)} \right)$$

$$= \frac{1}{N} \sum_{s=0}^{N-1} f(s) \delta_{ks} N = f(k).$$

$\square$

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 are 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.

Solution. The case (a). It follows directly from the definition of the Laplace transform that

$$\mathcal{L}(f)(s) = \int\limits_0^\infty e^{at} e^{-st} \, dt$$

$$= \int\limits_0^\infty e^{-(s-a)t} \, dt = \lim_{R\to\infty} \left(\frac{e^{-(s-a)R}}{-(s-a)} \right) - \frac{e^0}{-(s-a)} = \frac{1}{s-a}.$$

The case (b). Using the result of the above case and the linearity of improper integrals, we obtain

$$\mathcal{L}(f)(s) = c_1 \int\limits_0^\infty e^{a_1 t} e^{-st} \, dt + c_2 \int\limits_0^\infty e^{a_2 t} e^{-st} \, dt =$$
$$\frac{c_1}{s-a_1} + \frac{c_2}{s-a_2}.$$

The case (c). Since

$$\cos(bt) = \tfrac{1}{2}\left(e^{ibt} + e^{-ibt}\right),$$

the choice $c_1 = 1/2 = c_2$, $a_1 = ib$, $a_2 = -ib$ in the previous case gives

$$\mathcal{L}(f)(s) = \int\limits_0^\infty \left(\tfrac{1}{2}e^{ibt} + \tfrac{1}{2}e^{-ibt}\right) e^{-st} \, dt =$$
$$\frac{1}{2(s-ib)} + \frac{1}{2(s+ib)} = \frac{s}{s^2+b^2}.$$

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:

$$\mathcal{L}(t^k)(s) = \int_0^\infty t^k e^{-st} \, dt$$

$$= \left[-t^k \tfrac{1}{s} e^{-st} \right]_0^\infty + \frac{k}{s} \int_0^\infty t^{k-1} e^{-st} \, dt$$

$$= \cdots = \frac{k!}{s^k} \int_0^\infty e^{-st} \, dt = \frac{k!}{s^{k+1}}.$$

$\square$

7.D.2. 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.D.1(g).

$\bigcirc$

7.D.3. For $s > -1$, calculate the Laplace transform $\mathcal{L}(g)(s)$ of the function

$$g(t) = t\, e^{-t}.$$

A set X together with a mapping $d : X \times X \to \mathbb{R}$ 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.

The Euclidean distance in the vector spaces $\mathbb{R}^n$ satisfies the above three requirements.

If X is a vector space over $\mathbb{R}$ and $\| \ \| : X \to \mathbb{R}$ 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 λ,

(6) $\|x+y\| \leq \|x\| + \|y\|$,

then the function $\| \ \|$ is called a *norm* on X, and the space X is then a *normed vector space*.

The L_1 norm $\|\|_1$ and the L_2 norm $\|\|_2$ on functions, as well as the Euclidean norm on $\mathbb{R}^n$, satisfy these properties.

A norm always determines the metric $d(x,y) = \|x-y\|$, which is also the case with the Euclidean distance.

But not every metric can be defined by a norm in this way.

At the beginning of this chapter, we defined the distance between functions using the L_1–norm. In Euclidean vector spaces, it is the norm $\|x\|$, which is induced by the bilinear inner product by the relation $\|x\|^2 = \langle x, x \rangle$. Similarly, we work with the norm on unitary spaces. We obtained the L_2–norm on continuous functions in the same way.

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:

Further, for $s > 1$, calculate the Laplace transform $\mathcal{L}(h)(s)$ of the function

$$h(t) = t \sinh t.$$

○

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}	$\frac{1}{s-a}$
$t\,\mathrm{e}^{at}$	$\frac{1}{(s-a)^2}$
$t^n\,\mathrm{e}^{at}$	$\frac{n!}{(s-a)^{n+1}}$
$\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$
$\cos \omega t$	$\frac{s}{s^2+\omega^2}$
$\mathrm{e}^{at}\sin \omega t$	$\frac{\omega}{(s-a)^2+\omega^2}$
$\mathrm{e}^{at}(\cos \omega t + \frac{a}{\omega}\sin \omega t)$	$\frac{s}{(s-a)^2+\omega^2}$
$t\sin \omega t$	$\frac{2\omega s}{(s^2+\omega^2)^2}$
$\sin \omega t - \omega t \cos \omega t$	$\frac{2\omega^3}{(s^2+\omega^2)^2}$

7.D.4. Establish the 5th and 6th rows of the table above using Euler's formula $\mathrm{e}^{i\omega t} = \cos \omega t + i \sin \omega t$. ○

As expected, using the features of the Laplace transform allows us to find explicit solutions to some differential equations. By 7.H.8, 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.H.11. We return to this topic in Chapter 8.

E. 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 semimetrics). [2]

---

[2] 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.

Consider an arbitrary sequence of elements $x_0, x_1, \ldots$ in X. Suppose that for any fixed positive real number ε,

$$d(x_i, x_j) < \varepsilon$$

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:

Let $x_0, x_1, \ldots$ 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 ε, 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".

Definition The *open ε–neighbourhood* of an element x in a metric space X (or just ε–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 ε–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 ε–neighbourhood, we also talk about (open) ε–ball centered at x. In the case of a normed space, we can consider ε–balls centered at zero: along with x, ε–balls determine an ε–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.

We prove that a set is closed if and only if it contains all of its limit points:

7.E.1. The *discrete metric space* X is defined as the set X with the function $d : X \times X \to \mathbb{R}$

$$d(x,y) = \begin{cases} 1 & x \neq y \\ 0 & x = y. \end{cases}$$

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.

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.

$$d\big((x_1,y_1),(x_2,y_2)\big) = d_X(x_1,x_2) + d_Y(y_1,y_2).$$

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.E.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, $d(P_1, P_2) = n$,
   $P_1 = X_0, X_1, \ldots, X_{n+1} = P_2$ is the shortest seuquence of people, such that X_i knows X_{i+1} for $i = 0, \ldots n$.

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 $d(8,9) > d(8,6) + d(6,9)$.

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 ε–neighbourhood of x not intersecting with A. But in every ε–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 ε–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.

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 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, $i = 0, 1, \ldots$, 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:

Let W and Z be metric spaces. A mapping $f : W \to Z$ 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 ε, there is a positive real number δ 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.

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".

□

7.E.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 12.4.2). 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. □

7.E.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.F.10) 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*. □

7.E.5. Consider a space of integrable function on the interval $[a, b]$. Define the (L_1) distance of the functions f, g as

$$\|f, g\| = \int_a^b |f(x) - g(x)| \, \mathrm{d}x$$

. 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. □

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 $z = (z_1, \ldots, z_n)$, 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*: This is

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). 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 a Jensen inequality, see).

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

$$e^{v_k/p + w_k/q} \leq \tfrac{1}{p} \, e^{v_k} + \tfrac{1}{q} \, e^{w_k}.$$

By substitution, it follows immediately that

$$\frac{1}{XY} x_k y_k \leq \frac{1}{p} \left(\frac{x_k}{X} \right)^p + \frac{1}{q} \left(\frac{y_k}{Y} \right)^q.$$

Summing over $k = 1, \ldots, n$, gives

$$\frac{1}{XY} \sum_{i=1}^n x_i y_i \leq \frac{1}{pX^p} \sum_{i=1}^n x_i^p + \frac{1}{qY^q} \sum_{i=1}^n y_i^q$$

$$= \frac{1}{pX^p} X^p + \frac{1}{qY^q} Y^q = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying this inequality by XY finishes the proof. □

Now we can prove that $\| \, \|_p$ is indeed a norm:

7.E.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 $\|.\|_p : \mathbb{Q} \to \mathbb{R}$, $\|r\| \mapsto p^{-k}$. Show that it is a norm on $\mathbb{Q}$ as a vector space over $\mathbb{Q}$. It is called the *p-adic norm* ⃝

Solution. It is an exercise in elementary number theory. □

7.E.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) $d_1(X, Y) := |(X \cup Y) \smallsetminus (X \cap Y)|$,

(b) $d_2(X, Y) := \frac{|(X \cup Y) \smallsetminus (X \cap Y)|}{|X \cup Y|}$, 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)
$$(X \cup Z) \smallsetminus (X \cap Z) \subseteq \big((X \cup Y) \smallsetminus (X \cap Y)\big) \cup \big((Y \cup Z) \smallsetminus (Y \cap Z)\big)$$

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 $x \in (Y \cup Z) \smallsetminus (Y \cap Z)$, or $x \notin Y$ in which case $x \in (X \cup Y) \smallsetminus (X \cap Y)$. 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,
$$d_1(X, Z) = |(X \cup Z) \smallsetminus (X \cap Z)|$$
$$\leq \big|\big((X \cup Y) \smallsetminus (X \cap Y)\big) \cup \big((Y \cup Z) \smallsetminus (Y \cap Z)\big)\big|$$
$$\leq |(X \cup Y) \smallsetminus (X \cap Y)| + |(Y \cup Z) \smallsetminus (Y \cap Z)|$$
$$= d_1(X, Y) + d_1(Y, Z).$$

The case (b). Proceed similarly to the case of d_1. Denote by X' the complement of a set X. The equalities
$$(X \cup Y) \smallsetminus (X \cap Y) =$$
$$(X \cap Y' \cap Z) \cup (X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y \cap Z'),$$

MINKOWSKI INEQUALITY

For every $p > 1$ and all n–tuples of non-negative real numbers $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$,
$$\left(\sum_{i=1}^n (x_i + y_i)^p\right)^{1/p} \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} + \left(\sum_{i=1}^n y_i^p\right)^{1/p}.$$

To verify this inequality, we can use the following trick. By Hölder's inequality, (recall $p > 1$)
$$\sum_{i=1}^n x_i(x_i+y_i)^{p-1} \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} \cdot \left(\sum_{i=1}^n (x_i+y_i)^{(p-1)q}\right)^{1/q}$$
and
$$\sum_{i=1}^n y_i(x_i+y_i)^{p-1} \leq \left(\sum_{i=1}^n y_i^p\right)^{1/p} \cdot \left(\sum_{i=1}^n (x_i+y_i)^{(p-1)q}\right)^{1/q}.$$
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 $\|\ \|_p$ for all $p > 1$. The case $p = 1$ was considered earlier. We can also consider $p = \infty$ by setting
$$\|z\|_\infty = \max\{|z_i|, \ i = 1, \ldots, n\}.$$
This is a norm.

We notice that Hölder's inequality can, in the context of these norms, be written for all $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ 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 norms on suitable infinite-dimensional vector spaces as well. We begin with sequences.

The vector space ℓ_p, $p \geq 1$, is the set of all sequences of real or complex numbers $x_0, x_1, \ldots$ such that

$$(Y \cup Z) \smallsetminus (Y \cap Z) =$$
$$(X \cap Y \cap Z') \cup (X \cap Y' \cap Z) \cup (X' \cap Y \cap Z') \cup (X' \cap Y' \cap Z),$$

$$\big((X \cup Z) \smallsetminus (X \cap Z)\big) \cup \big(Y \smallsetminus (X \cup Z)\big) = (X \cap Y \cap Z') \cup$$
$$(X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y' \cap Z) \cup (X' \cap Y \cap Z'),$$

which, again, can be proved by listing several possibilities, imply a stronger form of (1), namely

$$\big((X \cup Z) \smallsetminus (X \cap Z)\big) \cup \big(Y \smallsetminus (X \cup Z)\big) \subseteq$$
$$\big((X \cup Y) \smallsetminus (X \cap Y)\big) \cup \big((Y \cup Z) \smallsetminus (Y \cap Z)\big).$$

Further, we invoke the inequality

$$\frac{|(X \cup Z) \smallsetminus (X \cap Z)|}{|X \cup Z|} \leq$$
$$\frac{\left|\big((X \cup Z) \smallsetminus (X \cap Z)\big) \cup \big(Y \smallsetminus (X \cup Z)\big)\right|}{\left|X \cup Z \cup \big(Y \smallsetminus (X \cup Z)\big)\right|}, \quad X \cup Z \neq \emptyset.$$

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

$$X \cup Z \cup \big(Y \smallsetminus (X \cup Z)\big) = X \cup Y \cup Z,$$

we obtain

$$\begin{aligned}
d_2(X, Z) &= \frac{|(X \cup Z) \smallsetminus (X \cap Z)|}{|X \cup Z|} \\
&\leq \frac{\left|\big((X \cup Z) \smallsetminus (X \cap Z)\big) \cup \big(Y \smallsetminus (X \cup Z)\big)\right|}{\left|X \cup Z \cup \big(Y \smallsetminus (X \cup Z)\big)\right|} \\
&\leq \frac{\left|\big((X \cup Y) \smallsetminus (X \cap Y)\big) \cup \big((Y \cup Z) \smallsetminus (Y \cap Z)\big)\right|}{|X \cup Y \cup Z|} \leq \\
&\leq \frac{|(X \cup Y) \smallsetminus (X \cap Y)| + |(Y \cup Z) \smallsetminus (Y \cap Z)|}{|X \cup Y \cup Z|} \\
&\leq \frac{|(X \cup Y) \smallsetminus (X \cap Y)|}{|X \cup Y|} + \frac{|(Y \cup Z) \smallsetminus (Y \cap Z)|}{|Y \cup Z|} = \\
&\qquad d_2(X, Y) + d_2(Y, Z),
\end{aligned}$$

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*.[3] □

7.E.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) $$f(t) := \frac{t}{1 + t}, \quad t \geq 0.$$

_____

[3]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.

$$\sum_{i=0}^{\infty} |x_i|^p < \infty.$$

If $x = (x_1, x_2 \ldots) \in \ell_p, \ p \geq 1$, then the norm is given by

$$\|x\|_p = \left(\sum_{i=0}^{\infty} |x_i|^p\right)^{1/p}$$

That $\|x\|_p$ is a norm follows immediately from the Minkowski inequality by letting $n \longrightarrow \infty$.

The vector space ℓ_∞, is the set of all bounded sequences of real or complex numbers $x_0, x_1, \ldots$.

If $x = (x_0, x_1, \ldots) \in \ell_\infty$, then its norm is given by

$$\|x\|_\infty = \sup\{|x_i|, \ i = 0, 1, 2, 3 \ldots\}$$

It is easily checked that this 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 $\| \ \|_1$. 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 Ξ with representatives ξ_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 of two functions $f(x)$ and $g(x)$ gives

$$\sum_{i=1}^n |f(\xi_i)| |g(\xi_i)| (x_i - x_{i-1}) =$$

$$= \sum_{i=1}^n |f(\xi_i)| (x_i - x_{i-1})^{1/p} |g(\xi_i)| (x_i - x_{i-1})^{1/q}$$

$$\leq \left(\sum_{i=1}^n |f(\xi_i)|^p (x_i - x_{i-1})\right)^{1/p} \cdot \left(\sum_{i=1}^n |g(\xi_i)|^q (x_i - x_{i-1})\right)^{1/q},$$

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.

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,

$$
\begin{aligned}
d(x, z) = \frac{|x - z|}{1 + |x - z|} &= \frac{|x - y + y - z|}{1 + |x - y + y - z|} \\
&\leq \frac{|x - y| + |y - z|}{1 + |x - y| + |y - z|} \\
&= \frac{|x - y|}{1 + |x - y| + |y - z|} + \frac{|y - z|}{1 + |x - y| + |y - z|} \\
&\leq \frac{|x - y|}{1 + |x - y|} + \frac{|y - z|}{1 + |y - z|} \\
&= d(x, y) + d(y, z), \quad x, y, z \in \mathbb{R}.
\end{aligned}
$$

$\square$

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.E.9. Determine the distance between the functions

$$f(x) = x, \quad g(x) = -\frac{x}{\sqrt{1+x^2}}, \quad x \in [1, 2]$$

as elements of the normed vector space $\mathcal{S}^0 [1, 2]$ of (piecewise) continuous functions on the interval $[1, 2]$ with norm

(a) $\| f \|_1 = \int_1^2 | f(x) | \, \mathrm{d}x$;

(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

$$
\begin{aligned}
\int_1^2 | f(x) - g(x) | \, \mathrm{d}x &= \int_1^2 x + \frac{x}{\sqrt{1+x^2}} \, \mathrm{d}x \\
&= \left(\frac{x^2}{2} + \sqrt{1 + x^2} \right)_1^2 = \frac{3}{2} + \sqrt{5} - \sqrt{2}.
\end{aligned}
$$

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

$$\left(x + \frac{x}{\sqrt{1+x^2}} \right)' = 1 + \frac{1}{\left(\sqrt{1+x^2}\right)^3} > 0, \quad x \in [1, 2],$$

it follows that $f - g$ is increasing, and so attains its maximum at the right end point of the interval when $x = 2$. So

$$\max_{x \in [1,2]} \left(x + \frac{x}{\sqrt{1+x^2}} \right) = 2 + \frac{2}{\sqrt{1+2^2}} = 2 + \frac{2}{\sqrt{5}}.$$

$\square$

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 $\| \ \|_p$ 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).

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_∞–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 $\mathbb{R}$.

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.

$\square$

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.E.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

$$\|f\| = \left(\int_a^b |f(x)|^2 + \alpha^2 |f'(x)|^2 \,\mathrm{d}\,x \right)^{1/2}$$

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 α.

Solution. The formula

$$\langle f, g \rangle \mapsto \int_a^b f(x)\overline{g(x)} + \alpha^2 f'(x)\overline{g'(x)} \,\mathrm{d}\,x$$

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 α 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.
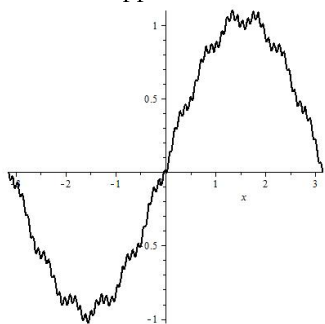
Let X be a metric space with metric d which is not complete. A metric space $\tilde{X}$ with metric $\tilde{d}$ such that $X \subset \tilde{X}$, d is the restriction of $\tilde{d}$ to the subset X and the closure $\bar{X}$ in $\tilde{X}$, is the whole space $\tilde{X}$ is called a *completion of the metric space* X.

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.

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 $x_i = x, i = 0, 1, \ldots$.

It is now easy to define the metric $\tilde{d}$. We put

$$\tilde{d}(\tilde{x}, \tilde{y}) = \lim_{i \to \infty} d(x_i, y_i)$$

for sequences $\tilde{x} = \{x_0, x_1, \ldots\}$ and $\tilde{y} = \{y_0, y_1, \ldots\}$.

First, we have to verify that this limit exists at all and is finite. Using the triangle inequality, and the fact that both the sequences $\tilde{x}$ and $\tilde{y}$ are Cauchy sequences, it follows that the considered sequence is also a Cauchy sequence of real numbers $d(x_i, y_i)$, so its limit exists.

If we select different representatives $\tilde{x} = \{x'_0, x'_1, \ldots\}$ and $\tilde{y} = \{y'_0, y'_1, \ldots\}$, then from the triangle inequality for the distance of real numbers (we need to consider the consequences for differences of distances) we see that

$$|d(x'_i, y'_i) - d(x_i, y_i)| \leq |d(x'_i, y'_i) - d(x'_i, y_i)| +$$
$$|d(x'_i, y_i) - d(x_i, y_i)|$$
$$\leq d(x_i, x'_i) + d(y_i, y'_i).$$

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

If $\alpha = 0$, the distance 0.162 is the usual L_2 distance. If $\alpha = 1$ the distance is 3.367.[4] $\square$

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.E.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. $\square$

F. Convergence

7.F.1. 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 $x_1, x_2, \ldots$ 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

---

[4]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.

the elements $\tilde{x}, \tilde{y}, \tilde{z}$, and we obtain

$$\tilde{d}(\tilde{x}, \tilde{z}) = \lim_{i \to \infty} d(x_i, z_i)$$
$$\leq \lim_{i \to \infty} d(x_i, y_i) + \lim_{i \to \infty} d(y_i, z_i)$$
$$= \tilde{d}(\tilde{x}, \tilde{y}) + \tilde{d}(\tilde{y}, \tilde{z}).$$

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 ε. Then the choice $y = x_n$ for one of those indices necessarily implies that the elements y and x_m are closer together than ε, 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 ε. 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 $\tilde{z} = \{z_0, z_1, \ldots\}$. 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:

$$d(z_m, z_n) = \tilde{d}(\tilde{z}_m, \tilde{z}_n)$$
$$\leq \tilde{d}(\tilde{z}_m, \tilde{x}_m) + \tilde{d}(\tilde{x}_m, \tilde{x}_n) + \tilde{d}(\tilde{x}_n, \tilde{z}_n)$$
$$\leq 1/m + \varepsilon/2 + 1/n \leq 2\frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon.$$

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.

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.F.2. Determine whether or not the sequence $\{x_n\}_{n\in\mathbb{N}}$ where

$$x_1 = 1, \quad x_n = 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{n}, \ n \in \mathbb{N} \smallsetminus \{1\},$$

is a Cauchy sequence in $\mathbb{R}$ using the standard metric.

Solution. Recall that

$$(1) \qquad \sum_{k=1}^{\infty} \frac{1}{k} = \infty, \quad \text{i.e.} \quad \sum_{k=m}^{\infty} \frac{1}{k} = \infty, \ m \in \mathbb{N}.$$

Therefore,

$$\lim_{n\to\infty} |x_n - x_m| = \sum_{k=m+1}^{\infty} \tfrac{1}{k} = \infty, \quad m \in \mathbb{N}.$$

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.F.3. Repeat the question from the previous problem with the metric d given by (cf. 7.E.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.F.4. 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) $\| f \|_p = \left(\int_{-1}^{1} |f(x)|^p \, \mathrm{d}x \right)^{1/p}$ 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

$$f_n(x) = 0, \ x \in [-1,0), \quad f_n(x) = 1, \ x \in \left(\tfrac{1}{n}, 1\right],$$

$$f_n(x) = nx, \ x \in \left[0, \tfrac{1}{n}\right].$$

A mapping $\varphi : X_1 \to X_2$ 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, $\iota_1 : X \to \tilde{X}_1$ and $\iota_2 : X \to \tilde{X}_2$, 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 φ. Thus if such a mapping φ 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 φ to the mapping $\tilde{\varphi} : \tilde{X}_1 \to \tilde{X}_2$ 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 $F : X \to X$ 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.

For every $m \geq n$, $m, n \in \mathbb{N}$, we compute the inequality

$$\left(\int\limits_{-1}^{1} |f_m(x) - f_n(x)|^p \, \mathrm{d}x \right)^{1/p} < \left(\int\limits_{0}^{1/n} 1 \, \mathrm{d}x \right)^{1/p} =$$
$$\left(\tfrac{1}{n} \right)^{1/p}.$$

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

$$f_n(x) = 0, \ x \in [-1, 0], \quad f_n(x) = 1, \ x \in [\varepsilon, 1]$$

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 δ, since f is continuous. Thus $f \neq 1$ on some bounded interval containing y. Therefore f must satisfy

$$f(x) = 0, \ x \in [-1, 0], \quad f(x) = 1, \ x \in [\varepsilon, 1]$$

for an arbitrarily small $\varepsilon > 0$. Thus, necessarily,

$$f(x) = 0, \ x \in [-1, 0], \quad f(x) = 1, \ x \in (0, 1].$$

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

$$(1) \qquad \max_{x \in [-1, 1]} |f_m(x) - f_n(x)| < \frac{\varepsilon}{2}, \quad m, n \geq n(\varepsilon).$$

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

$$f(x) := \lim_{n \to \infty} f_n(x), \quad x \in [-1, 1].$$

Letting $m \to \infty$ in (1), we obtain

$$\max_{x \in [-1, 1]} |f(x) - f_n(x)| \leq \tfrac{\varepsilon}{2} < \varepsilon, \quad n \geq n(\varepsilon).$$

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

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, $i = 0, 1, \ldots$

$$z_1 = F(z_0), \ z_2 = F(z_1), \ldots, z_{i+1} = F(z_i), \ldots$$

From the assumptions, we have

$$d(z_{i+1}, z_i) = d(F(z_i), F(z_{i-1}))$$
$$\leq C \, d(z_i, z_{i-1}) \leq \cdots \leq C^i \, d(z_1, z_0).$$

The triangle inequality then implies that for all natural numbers j,

$$d(z_{i+j}, z_i) \leq \sum_{k=1}^{j} d(z_{i+k}, z_{i+k-1})$$

$$\leq \sum_{k=1}^{j} C^{i+k-1} \, d(z_1, z_0) = C^i \, d(z_1, z_0) \sum_{k=1}^{j} C^{k-1}$$

$$\leq C^i \, d(z_1, z_0) \sum_{k=1}^{\infty} C^{k-1} = \frac{C^i}{1 - C} \, d(z_1, z_0).$$

Now, since $0 \leq C < 1$, $\lim_{n \to \infty} C^n = 0$, so for every positive (no matter how small) ε, the right-hand expression is surely less than ε 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 thre 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,

$$F(z) = F(\lim_{n \to \infty} z_n) = \lim_{n \to \infty} F(z_n) = z.$$

This finishes the proof. $\square$

The next two theorems extend the intutive 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

$$\operatorname{diam} A = \sup_{x, y \in A} d(x, y)$$

is called the *diameter of the set A*. The set A is said to be *bounded* if and only if $\operatorname{diam} A < \infty$.

Theorem. *If $A_1 \supset A_2 \supset \cdots \supset A_i \supset \ldots$ is a non-increasing sequence of non-empty closed subsets in a complete metric space X and if $\operatorname{diam} A_i \to 0$, then there is exactly one point $x \in X$ belonging to the intersection of all the sets A_i.*[2]

---

[2]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,

functions on any closed bounded interval $[a, b]$ or on the space $\mathcal{C}_c$ of continuous functions with compact support.

7.F.5. Prove that the metric space ℓ_2 is complete.

Solution. Recall that ℓ_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 ℓ_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 ℓ_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 $n \to \infty$ in ℓ_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

$$\sum_{k=1}^{\infty} \left(x_m^k - x_n^k\right)^2 < \varepsilon, \quad m, n \geq n(\varepsilon),\, m, n \in \mathbb{N}.$$

In particular,

$$\sum_{k=1}^{l} \left(x_m^k - x_n^k\right)^2 < \varepsilon, \quad m, n \geq n(\varepsilon),\, m, n, l \in \mathbb{N},$$

whence, letting $m \to \infty$,

$$\sum_{k=1}^{l} \left(z_k - x_n^k\right)^2 \leq \varepsilon, \quad n \geq n(\varepsilon),\, n, l \in \mathbb{N},$$

i.e. (this time $l \to \infty$)

$$(1) \qquad \sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 \leq \varepsilon, \quad n \geq n(\varepsilon),\, n \in \mathbb{N}.$$

Especially,

$$\sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 < \infty, \quad n \geq n(\varepsilon),\, n \in \mathbb{N}$$

and, at the same time,

$$\sum_{k=1}^{\infty} \left(x_n^k\right)^2 < \infty, \quad n \in \mathbb{N},$$

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)

$$\sum_{k=1}^{\infty} \left(z_k x_n^k\right) \leq \sqrt{\sum_{k=1}^{\infty} z_k^2} \cdot \sqrt{\sum_{k=1}^{\infty} \left(x_n^k\right)^2}, \quad n \in \mathbb{N}$$

and

PROOF. Select one point x_i for each set A_i. Since $\operatorname{diam} A_i \to 0$, for every positive real number ε, we can find an index $n(\varepsilon)$ such that for all A_i with indices $i \geq n(\varepsilon)$, their diameters are less than ε. 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 $\operatorname{diam} A_i \to 0$, 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.*[3]

PROOF. Suppose X contains a system of dense open sets A_i, $i = 1, 2 \ldots$. It is required to show that the set $A = \cap_{i=1}^{\infty} A_i$ has a non-empty intersection with any open set $U \subset X$. Proceed inductively, invoking the previous theorem.

Since A_1 is dense in X, there is a point $z_1 \in A_1 \cap U$. Let U_1 be an open ball, centre z_1, of positive radius ε_1, such that its closure B_1 is contained in U.

Suppose the points z_i and their open ε_i–neighbourhoods U_i are already chosen for $i = 1, \ldots, n$ with $z_i \in A_i \cap U_i$.

Since the set A_{n+1} is open and dense in X, there is a point $z_{n+1} \in A_{n+1} \cap \bar{U}_n$; however, since $A_{n+1} \cap U_n$ is open, the point z_{n+1} belongs to it together with a sufficiently small ε_{n+1}-neighbourhood U_{n+1}.

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 ε_1-neighbourhood U_1 (for sufficiently small ε_1) of the point z_1 is contained in A_1 as well. Denote the closure of this ε_1–ball U_1 by B_1.

Further, suppose that the points z_i and their open ε_i–neighbourhoods U_i are already chosen for $i = 1, \ldots, n$. Since the set A_{n+1} is open and dense in X, there is a point $z_{n+1} \in A_{n+1} \cap \bar{U}_n$; however, since $A_{n+1} \cap U_n$ is open, the point z_{n+1} belongs to it together with a sufficiently small ε_{n+1}-neighbourhood U_{n+1}.

Then, the closures surely satisfy $B_{n+1} = \bar{U}_{n+1} \subset \bar{U}_n$, and so the closed set B_{n+1} is contained in $A_{n+1} \cap \bar{U}_n$. Moreover, we can assume that $\varepsilon_n \leq 1/n$.

If we proceed in this inductive way from the original point z_1 and the set B_1, we obtain a non-decreasing sequence of non-empty closed sets B_n whose diameter approaches

the new abstract approach to fundamentals of Mathematics caused fierce objections. It also lead to the severe internal crises of Mathematics in the beginning of the 20th century. This part of the history of Mathematics is fascinating.

[3]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.

$$\sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 = \sum_{k=1}^{\infty} \left(z_k^2 - 2z_k x_n^k + \left(x_n^k\right)^2\right), \quad n \in \mathbb{N}.$$

Hence

$$\sum_{k=1}^{\infty} z_k^2 < \infty.$$

It is proved that $z \in \ell_2$. The fact that $\{x_n\}$ converges for $n \to \infty$ to z in ℓ_2 follows from (1). $\qquad\square$

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.F.6. 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 sequnce 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 $n \to \infty$, then

$$\int_{-\infty}^{\infty} |f_n(x)|^q dx \longrightarrow 0.$$

So $\| f_n \|_q \longrightarrow 0$, and the sequence converges to the zero function. Similarly if $0 < p < q$ and if $n \to \infty$, then

$$\int_{-\infty}^{\infty} |f_n(x)|^q dx \longrightarrow \infty.$$

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 $n \longrightarrow \infty$ we get $\| f_n \|_p \longrightarrow 2^{1/p}$. 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 $q \geq p$. $\qquad\square$

The next problem deals with the extremely useful Banach fixed point theorem, showing the necessity of all the requirements in Theorem 7.3.9.

zero. Therefore, there is a point z common to all of these sets. That is,

$$z \in \cap_{i=1}^{\infty} \bar{U}_i = \cap_{i=1}^{\infty} B_i \subset \cap_{i=1}^{\infty} A_n \cap U,$$

which is the statement to be proved. $\qquad\square$

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 ε–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 ε–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 R, our definition reveals the compact subsets discussed there and we would also like to come to useful properties as we did for real numbers in the paragraphs 5.2.7–5.2.8. It is suprisingly easy to see, that the continuous functions behave similarly on compact sets in general:

Theorem. *Let $f : X \to Y$ be a continuous mapping between metric spaces. Then the images of compact sets are compact.*

PROOF. Recall that any convergent sequence of points $x_i \to x$ in X is mapped onto the convergent sequence $f(x_i) \to f(x)$ in Y. Thus, the statement follows immediately from our definition of the compactness via convergent subsequences. $\qquad\square$

In particular we obtain the most useful consequence on the minima and maxima of continuous functions on compact subsets:

Corollary. *Let $f : X \to \mathbb{R}$ be a real function defined on a compact metric space. Then there are the points x_0 and y_0 in*

7.F.7. Show that the mapping $f : \langle 0, \infty \rangle \to \langle 0, \infty \rangle$ given by

$$f(x) = x + \mathrm{e}^{-x}$$

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 $\mathrm{e}^{-x} < \mathrm{e}^{-y}$ and

$$|f(x) - f(y)| = |x - y + \mathrm{e}^{-x} - \mathrm{e}^{-y}| < |x - y|.$$

Finally, $f(x) = x$ implies $\mathrm{e}^{-x} = 0$ which is impossible. $\quad\square$

G. Topology

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.G.1. We have already seen the discrete metric space $X \neq \emptyset$ with the metric $d : X \times X \to \mathbb{R}$ defined by the formula

$$d(x, y) := 1, \; x \neq y, \quad d(x, y) := 0, \; x = y \,.$$

(a) Decide whether (X, d) is complete.

(b) Describe all open, closed, and bounded sets in (X, d).

(c) Describe the interior, boundary, limit, and isolated points of an arbitrary set in (X, d).

(d) Describe all compact sets in (X, d).

Solution. The case (a) was essentially dealt with in 7.F.1. 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 $x_n = x_{n+1} = \cdots$ (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,

---

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 the supremum and the infimum of the bounded and closed image). $\quad\square$

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*[4]

$$\mathrm{dist}(x, B) = \inf_{y \in B}\{d(x, y)\}.$$

We say that a metric space X is *totally bounded* if for every positive real number ε, there is a finite set A such that

$$\mathrm{dist}(x, A) < \varepsilon$$

for all points $x \in X$. We call such an A an *ε-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 ε from the definition of total boundedness, then the distance $d(x, y)$ of two points can always be bounded by the sum of $\mathrm{dist}(x, A)$, $\mathrm{dist}(y, A)$, and $\mathrm{diam}\, A$, 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 $(1) \implies (3) \implies (2) \implies (1)$.

$(1) \implies (3)$. Assume 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, $x_i \to x$. 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 ε-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

---

[4]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 $\mathrm{dist}(A, B) = \max\{\sup\{\mathrm{dist}(x, B), x \in A\}, \sup\{\mathrm{dist}(y, A), y \in B\}\}$.

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 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). □

7.G.2. In the metric space $\mathcal{S}[-1, 1]$ with metric given by the norm $\|\cdot\|_\infty$, consider the sets

$$A = \{f \in \mathcal{S}[-1, 1]; \ f(0) \in (0, 2)\},$$

$$B = \{f \in \mathcal{S}[-1, 1]; \ \int_{-1}^{1} f(x) \, \mathrm{d}x = 0\}.$$

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. □

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,

$$(1) \qquad \bigcap_{n \in \mathbb{N}} A_n \neq \emptyset.$$

7.G.3. Verify that the additional condition in the theorem

$$(1) \qquad \lim_{n \to \infty} \sup \{d(x, y); \ x, y \in A_n\} = 0.$$

cannot be omitted.

almost obvious claim – look at the definition of ε-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.

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_α, $\alpha \in I$, of X, which does not contain any finite covering. Choose a sequence of positive real numbers $\varepsilon_k \to 0$ and consider the finite ε_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 ε_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 the sets U_α. Call it C_1 and notice that $\operatorname{diam} C_1 = 2\varepsilon_1$.

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_α, we call it C_2. This way, we inductively construct a sequence of sets C_k satisfying $C_{k+1} \subset C_k$, $\operatorname{diam} C_k \leq 2\varepsilon_k$, $\varepsilon_k \to 0$, and none of them can be covered by a finite number of the open sets U_α.

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_{α_0} containing x and containing also some δ-neighbourhood $B_\delta(x)$. But now, if $\operatorname{diam} C_k \leq 2\varepsilon_k < \delta$, 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 $C_n = \overline{\{x_k; \ k \geq n\}}$. 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_α, $\alpha \in I$, such that each its finite subsystem $D_{\alpha_1}, \ldots, D_{\alpha_k}$ has non-empty intersection. Then also*

$$\cap_{\alpha \in I} D_\alpha \neq \emptyset.$$

This simple lemma is proved by contradiction, again. If the latter intersection is empty, then

$$X = X \setminus (\cap_{\alpha \in I} D_\alpha) = \cup_{\alpha \in I} (X \setminus D_\alpha) = \cup_{\alpha \in I} V_\alpha,$$

where $V_\alpha = X \setminus D_\alpha$ are open sets. Thus, there must be a finite number of them, $\{V_{\alpha_1}, \ldots, V_{\alpha_n}\}$, covering X too. Thus, we obtain

$$X = \cup_{i=1}^{n} V_{\alpha_i} = \cup_{i=1}^{n} (X \setminus D_{\alpha_i}) = X \setminus (\cap_{i=1}^{n} D_{\alpha_i}).$$

This is a contradiction with our assumptions on D_α and the lemma is proved.

Now, let $x \in \cap_{n=1}^{\infty} C_n$. 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. □

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

$$d(m,n) = 1 + \tfrac{1}{m+n}, \ m \neq n, \quad d(m,n) = 0, \ m = n.$$

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 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

$$A_n := \left\{ m \in \mathbb{N}; \ d(m,n) \leq 1 + \tfrac{1}{2n} \right\}, \quad n \in \mathbb{N}.$$

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

$$\lim_{n\to\infty} \sup \left\{ d(x,y); \ x, y \in A_n \right\}$$
$$= \lim_{n\to\infty} \left(1 + \frac{1}{2n+1} \right) = 1 \neq 0.$$

$\square$

7.G.4. Determine whether the set (known as the Hilbert cube)

$$A = \left\{ \{x_n\}_{n\in\mathbb{N}} \in \ell_2; \ |x_n| \leq \tfrac{1}{n}, \ n \in \mathbb{N} \right\}$$

is compact in ℓ_2. Then determine the compactness of the set

$$B = \left\{ \{x_n\}_{n\in\mathbb{N}} \in \ell_\infty; \ |x_n| < \tfrac{1}{n}, \ n \in \mathbb{N} \right\}$$

in the space ℓ_∞.

Solution. The space ℓ_2 is complete (see 7.F.5). Every closed subset of a complete metric space defines a complete metric space. The set A is evidently closed in l_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 ε-net of A for any given $\varepsilon > 0$: Begin with the well-known series

$$\sum_{k=1}^{\infty} \tfrac{1}{k^2} = \tfrac{\pi^2}{6}$$

(see (1)).

For every $\varepsilon > 0$, there is an $n(\varepsilon) \in \mathbb{N}$ satisfying

$$\sqrt{\sum_{k=n(\varepsilon)+1}^{\infty} \tfrac{1}{k^2}} < \tfrac{\varepsilon}{2}.$$

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. If fact, many ideas understood for the functions of one real variable generalize naturally.

In particular, every continuous function $f : X \to \mathbb{R}$ on a compact set X is bounded and achieves its maximum and minimum. Indeed, 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 covergent 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.

A mapping $f : X \to Y$ 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 that this requirement on the uniform continuity of f is equivalent to the condition that for each pair of sequences x_k and y_k in X, $d_X(x_k, y_k) \to 0$ implies $d_Y(f(x_k), f(y_k)) \to 0$.

This observation leads to the following generalization of the behavior of real functions:

Lemma. *Each continuous mapping $f : X \to Y$ 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 $d(x_k, y_k) \to 0$.

Since X is compact, there is a subsequence of x_k converging to some point $x \in X$ and so we may assume $x_k \to x$, without loss of generality. Now, $d_X(x, y_k) \leq d_X(x, x_k) + d_X(x_k, y_k) \to 0$ and so $\lim_{k\to\infty} y_k = x$, too.

From each of the intervals $[-1/n, 1/n]$ for $n \in \{1, \ldots, n(\varepsilon)\}$, choose finitely many points $x_1^n, \ldots, x_{m(n)}^n$ so that for any $x \in [-1/n, 1/n]$ that

$$\min_{j \in \{1, \ldots, m(n)\}} |x - x_j^n| < \frac{\varepsilon}{\sqrt{5^n}}.$$

Consider such sequences $\{y_n\}_{n \in \mathbb{N}}$ from l_2 whose terms with indices $n > n(\varepsilon)$ are zero, and at the same time,

$$y_1 \in \left\{ x_1^1, \ldots, x_{m(1)}^1 \right\}, \ldots, y_{n(\varepsilon)} \in \left\{ x_1^{n(\varepsilon)}, \ldots, x_{m(n(\varepsilon))}^{n(\varepsilon)} \right\}.$$

There are only finitely many such sequences and they create the desired ε-net for A: let $x_n \in l_2$ is arbitrary. According to our choice of the sequences y_n, there is y_n such that

$$
\begin{aligned}
d(x_n, y_n) &= \sqrt{\sum_{k=1}^{\infty} (x_k - y_k)^2} \\
&\leq \sqrt{\sum_{k=1}^{n(\varepsilon)} (x_k - y_k)^2} + \sqrt{\sum_{k=n(\varepsilon)+1}^{\infty} x_k^2} \\
&\leq \sqrt{\frac{\varepsilon^2}{5} + \frac{\varepsilon^2}{5^2} + \cdots + \frac{\varepsilon^2}{5^{n(\varepsilon)}}} + \frac{\varepsilon}{2} \\
&< \varepsilon \cdot \sqrt{\frac{1}{1 - \frac{1}{5}} - 1} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}
$$

Since $\varepsilon > 0$ is arbitrary, the set A is totally bounded, which implies compactness.

The closure of the set B is

$$\overline{B} = \left\{ \{x_n\}_{n \in \mathbb{N}} \in l_\infty; \, |x_n| \leq \frac{1}{n}, \, n \in \mathbb{N} \right\}.$$

Hence B is not closed, and so it is not compact. The set $\overline{B}$ 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. □

7.G.5. Prove that on each metric space X, the given metric d is a continuous function $X \times X \to \mathbb{R}$. ○

7.G.6. 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 α 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.G.4). 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. □

Next, notice that the metric $d_Y : Y \times Y \to \mathbb{R}$ is always a continuous function (cf. the problem 7.E.1 in the other column). But then the continuity of f ensures

$$\lim_{k \to \infty} d_Y(f(x_k), f(y_k)) = d_Y(f(x), f(x)) = 0.$$

By the latter observation, this is equivalent to the uniform continuity of f. □

A very useful variation on the theme of continuity is the following definition.

LIPSCHITZ CONTINUITY

A function $f : X \to Y$ 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 $f : X \to Y$ between matric 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 $\| \, \|_\infty$ 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 $\| \, \|_\infty$ is compact if and only if it is bounded, closed, and equicontinuous.*[5]

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.

---

[5] 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.

Given $\varepsilon > 0$, consider the corresponding ε-net $(f_1, f_2, \ldots, f_k) \subset M$ 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 δ_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 δ to be the minimum of the finite many $\delta_i, i = 1, ..., k$. Then the same equality holds for all f_i in the ε-net. But now, considering an arbitrary function $f \in M$, there is a function f_j in our ε-net with $\|f - f_j\|$ and so $d_X(x, y) < \delta$ implies that $|f(x) - f(y)|$ is at most

$$|f(x) - f_j(x)| + |f_j(x) - f_j(y)| + |f_j(y) - f(y)| \le 3\varepsilon,$$

and the equicontinuity has been proved.

Conversely, suppose that M is a bounded, closed, and equicontinuous subset of $C(X)$, with X as a compact metric space.

First we show that M is complete. This was shown in the case when X is a closed bounded interval of reals in the problem 7.F.4. Exploit the equicontinuity to see that the limit function f is again continuous. The same argument works in general.

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 $A = \{a_1, a_2, \ldots\}$ as a sequence.

Choose a subsequence of functions $f_{1j}, j = 1, 2, \ldots$ 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 $\|\ \|_\infty$ 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).

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 δ_ε. Let $A_\varepsilon \subset A$ be subset forming a δ_ε-net. This is a finite set and so there must be an $n \in \mathbb{N}$ such that for all $i, j \ge n$ 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

$$|g_i(x) - g_i(t)| + |g_i(t) - g_j(t)| + |g_j(t) - g_j(x)| \le 3\varepsilon.$$

Thus, the sequence g_k is a Cauchy sequence in $C(X)$, and so M is compact. $\qquad\square$

H. Additional exercises to the whole chapter

7.H.1. Expand the function $\sin^2(x)$ on the interval $[-\pi, \pi]$ into a Fourier series. $\bigcirc$

7.H.2. Expand the function $\cos^2(x)$ on the interval $[-\pi, \pi]$ into a Fourier series. $\bigcirc$

7.H.3. 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

$$(1) \qquad x = \pi - 2 \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}, \quad x \in (0, 2\pi),$$

$$(2) \qquad x^2 = \frac{4\pi^2}{3} + 4 \sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2} - 4\pi \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}, \quad x \in (0, 2\pi),$$

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:

$$\sum_{n=1}^{\infty} \frac{\sin(nx)}{n^3} = \sum_{n=1}^{\infty} \left[\frac{\sin(ny)}{n^3} \right]_0^x = \int_0^x \sum_{n=1}^{\infty} \frac{\cos(ny)}{n^2} \, dy$$

$$= \int_0^x \frac{3y^2 - 6\pi y + 2\pi^2}{12} \, dy = \frac{x^3 - 3\pi x^2 + 2\pi^2 x}{12}, \quad x \in [0, 2\pi].$$

In fact, every Fourier series may be integrated term by term. Further integration gives

$$\sum_{n=1}^{\infty} \frac{1 - \cos(nx)}{n^4} = \sum_{n=1}^{\infty} \left[-\frac{\cos(ny)}{n^4} \right]_0^x = \int_0^x \sum_{n=1}^{\infty} \frac{\sin(ny)}{n^3} \, dy$$

$$= \int_0^x \frac{y^3 - 3\pi y^2 + 2\pi^2 y}{12} \, dy = \frac{x^4 - 4\pi x^3 + 4\pi^2 x^2}{48}, \quad x \in [0, 2\pi].$$

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

$$(3) \qquad \sum_{n=1}^{\infty} \frac{2}{(2n-1)^4} = \frac{\pi^4}{48}.$$

From the expression

$$\sum_{n=1}^{\infty} \tfrac{1}{n^4} = \sum_{n=1}^{\infty} \tfrac{1}{(2n)^4} + \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} = \tfrac{1}{16} \sum_{n=1}^{\infty} \tfrac{1}{n^4} + \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4},$$

it follows that

$$\sum_{n=1}^{\infty} \tfrac{1}{n^4} = \tfrac{16}{15} \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} = \tfrac{16}{15} \cdot \tfrac{1}{2} \cdot \tfrac{\pi^4}{48} = \tfrac{\pi^4}{90},$$

thereby having summed up the first series. As for the second one,

$$\begin{aligned}
\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n^4} &= \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} - \sum_{n=1}^{\infty} \tfrac{1}{(2n)^4} = \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} - \tfrac{1}{16} \sum_{n=1}^{\infty} \tfrac{1}{n^4} \\
&= \tfrac{1}{2} \cdot \tfrac{\pi^4}{48} - \tfrac{1}{16} \cdot \tfrac{\pi^4}{90} = \tfrac{7\pi^4}{720}.
\end{aligned}$$

One can proceed similarly to sum the series

$$\sum_{n=1}^{\infty} \tfrac{1}{n^{2k}}, \qquad \sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n^{2k}}$$

for other $k \in \mathbb{N}$.

It is natural to ask for the value of the series $\sum_{n=1}^{\infty} \tfrac{1}{n^3}$. 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

$$\sum_{n=1}^{\infty} \tfrac{\cos(nx)}{n} = -\ln\left(2\sin\tfrac{x}{2}\right), \qquad x \in (0, 2\pi),$$

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 $x \to 0+$ in the limit, we get the series $\sum_{n=1}^{\infty} \tfrac{1}{n^3}$. 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. [5]

$\square$

7.H.4. In problem 7.45 there occurs the following integral:

$$f(t) = \tfrac{1}{2\pi} \int_{-\infty}^{\infty} \tfrac{\sin\omega}{\omega} e^{i\omega t} \, d\omega$$

There, the integral was evaluated by converting the complex exponential to real trigonometric functions. Evaluate it by converting the real function $\sin(\omega)$ to its complex exponential form. $\bigcirc$

7.H.5. Determine the convolution of the functions f_1 and f_2, where

$$f_1 = \begin{cases} 1 & \text{for } x \in [-1, 0] \\ 0 & \text{otherwise} \end{cases}$$

$$f_2 = \begin{cases} x & \text{for } x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

$\bigcirc$

7.H.6. Determine the function f whose Fourier transform is the function

$$\tilde{f}(\omega) = \tfrac{1}{\sqrt{2\pi}} \tfrac{\sin\omega}{\omega}, \qquad \omega \neq 0.$$

Solution. We might have noticed that the sinc function appeared as the image of the characteristic function h_Ω of the interval $(-\Omega, \Omega)$ in one of the previous problems:

$$\tilde{h}_\Omega(\omega) = \tfrac{2\Omega}{\sqrt{2\pi}} \operatorname{sinc}(\omega\Omega).$$

---

[5]The function $\zeta(p) = \sum_{n=1}^{\infty} \tfrac{1}{n^p}$ is called the Riemann zeta function.
EXPAND THE FOOTNOTE!

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

$$f(t) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \frac{\sin \omega}{\omega} \, \mathrm{e}^{i\omega t} \, d\omega$$

$$= \frac{1}{2\pi} \left(\int\limits_{-\infty}^{0} \frac{\sin \omega}{\omega} \, \mathrm{e}^{i\omega t} \, d\omega + \int\limits_{0}^{\infty} \frac{\sin \omega}{\omega} \, \mathrm{e}^{i\omega t} \, d\omega \right).$$

Substitute $-\omega$ for ω in the first integral, to obtain

$$f(t) = \frac{1}{2\pi} \left(\int\limits_{0}^{\infty} \frac{\sin \omega}{\omega} \, \mathrm{e}^{-i\omega t} \, d\omega + \int\limits_{0}^{\infty} \frac{\sin \omega}{\omega} \, \mathrm{e}^{i\omega t} \, d\omega \right)$$

$$= \frac{1}{2\pi} \int\limits_{0}^{\infty} \frac{\sin \omega}{\omega} 2(\mathrm{e}^{-i\omega t} + \mathrm{e}^{i\omega t}) \, d\omega = \frac{1}{\pi} \int\limits_{0}^{\infty} \frac{\sin \omega}{\omega} \cos (\omega t) \, d\omega.$$

Continue via trigonometric identities to obtain

$$f(t) = \frac{1}{2\pi} \left(\int\limits_{0}^{\infty} \frac{\sin(\omega(1+t))}{\omega} \, d\omega + \int\limits_{0}^{\infty} \frac{\sin(\omega(1-t))}{\omega} \, d\omega \right).$$

The substitutions $u = \omega(1 + t)$, $v = \omega(1 - t)$ then give

$$f(t) = \frac{1}{2\pi} \left(\int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u - \int\limits_{0}^{\infty} \frac{\sin v}{v} \, \mathrm{d}\, v \right) = 0, \quad t > 1;$$

$$f(t) = \frac{1}{2\pi} \left(\int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u + \int\limits_{0}^{\infty} \frac{\sin v}{v} \, \mathrm{d}\, v \right) = \frac{1}{\pi} \int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u, \, t \in (-1, 1);$$

$$f(t) = \frac{1}{2\pi} \left(-\int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u + \int\limits_{0}^{\infty} \frac{\sin v}{v} \, \mathrm{d}\, v \right) = 0, \quad t < -1.$$

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

$$\int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u = \frac{\pi}{2}.$$

Alternatively, we can 'guess' that the constant is one, i.e.

$$g(t) = 1, \quad |t| < 1; \qquad g(t) = 0, \quad |t| > 1$$

and compute

$$\mathcal{F}(g)(\omega) = \frac{1}{\sqrt{2\pi}} \int\limits_{-1}^{1} \mathrm{e}^{-i\omega t} \, \mathrm{d}\, t = \frac{2}{\sqrt{2\pi}} \int\limits_{0}^{1} \cos (\omega t) \, \mathrm{d}\, t = \frac{2}{\sqrt{2\pi}} \frac{\sin \omega}{\omega}.$$

So $f(0) = g(0)/2 = 1/2$, which also establishes

$$\int\limits_{0}^{\infty} \frac{\sin u}{u} \, \mathrm{d}\, u = \frac{\pi}{2}.$$

$\square$

7.H.7. Using the relation

(1) $\qquad \mathcal{L}(f')(s) = s\mathcal{L}(f)(s) - \lim\limits_{t \to 0+} f(t),$

derive the Laplace transforms of both the functions $y = \cos t$ and $y = \sin t$.

Solution. Notice first that from (1), it follows that

$$\mathcal{L}(f'')(s) = s\mathcal{L}(f')(s) - \lim\limits_{t \to 0+} f'(t) = s \left(s\mathcal{L}(f)(s) - \lim\limits_{t \to 0+} f(t) \right) - \lim\limits_{t \to 0+} f'(t) =$$
$$s^2 \mathcal{L}(f)(s) - s \lim\limits_{t \to 0+} f(t) - \lim\limits_{t \to 0+} f'(t).$$

Therefore,

$$-\mathcal{L}(\sin t)(s) = \mathcal{L}(-\sin t)(s) = \mathcal{L}((\sin t)'')(s) = s^2 \mathcal{L}(\sin t)(s) - s \lim\limits_{t \to 0+} \sin t - \lim\limits_{t \to 0+} \cos t = s^2 \mathcal{L}(\sin t)(s) - 1,$$

whence we get

$$-\mathcal{L}\left(\sin t\right)(s) = s^2 \mathcal{L}\left(\sin t\right)(s) - 1, \quad \text{i. e.} \quad \mathcal{L}\left(\sin t\right)(s) = \frac{1}{s^2+1}.$$

Now, invoking (1), we determine

$$\mathcal{L}\left(\cos t\right)(s) = \mathcal{L}\left((\sin t)'\right)(s) = s\,\frac{1}{s^2+1} - \lim_{t \to 0+}\sin t = \frac{s}{s^2+1}.$$

$\square$

7.H.8. 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

$$\mathcal{L}(y')(s) = s\mathcal{L}(y)(s) - y(0)$$
$$\mathcal{L}(y'')(s) = s^2 \mathcal{L}(y) - sy(0) - y'(0)$$

and the claim is verified by induction. $\square$

7.H.9. 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.

$$\mathcal{L}(H(t))(s) = \int_0^\infty H(t)\,\mathrm{e}^{-st}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}t$$

$$= \left[-\frac{\mathrm{e}^{-st}}{s}\right]_0^\infty = -\frac{1}{s}(0-1) = \frac{1}{s},$$

$$\mathcal{L}(H_a(t))(s) = \mathcal{L}(H(t-a))(s)$$

$$= \int_0^\infty H(t-a)\,\mathrm{e}^{-st}\,\mathrm{d}t = \int_a^\infty \mathrm{e}^{-st}\,\mathrm{d}t$$

$$= \int_0^\infty \mathrm{e}^{-s(t+a)}\,\mathrm{d}t = \mathrm{e}^{-as}\,\mathcal{L}(H(t))(s) = \frac{\mathrm{e}^{-as}}{s}.$$

$\square$

7.H.10. Show that for real a,

(1) $\qquad \mathcal{L}(f(t) \cdot H_a(t))(s) = \mathrm{e}^{-as}\,\mathcal{L}(f(t+a))(s)$

Solution.

$$\mathcal{L}(f(t)H_a(t))(s) = \int_0^\infty f(t)H(t-a)\,\mathrm{e}^{-st}\,\mathrm{d}t = \int_a^\infty f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t$$

$$= \int_0^\infty f(t+a)\,\mathrm{e}^{-s(t+a)}\,\mathrm{d}t = \mathrm{e}^{-as}\int_0^\infty f(t+a)\,\mathrm{e}^{-st}\,\mathrm{d}t$$

$$= \mathrm{e}^{-as}\,\mathcal{L}(f(t+a))(s).$$

$\square$

7.H.11. 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.H.8:

$$s^2 \mathcal{L}(y)(s) + 4\mathcal{L}(y)(s) = \mathcal{L}(\sin 2t)(s)$$

Now, by 7.D.1(d)

$$\mathcal{L}(\sin 2t)(s) = \frac{2}{s^2 + 4}.$$

It follows that

$$\mathcal{L}(y)(s) = \frac{2}{(s^2 + 4)^2}.$$

More details ?   The inverse transform then gives

$$y(t) = \tfrac{1}{8} \sin 2t - \tfrac{1}{4} t \cos 2t.$$

□

7.H.12. Find a function $y(t)$ satisfying the differential equation and the initial conditions:

$$y''(t) + 4y(t) = f(t), \ y(0) = 0, \ y'(0) = -1,$$

where $f(t)$ is the piecewise continuous function

$$f(t) = \begin{cases} \cos(2t) & \text{for } 0 \le t < \pi, \\ 0 & \text{for } t \ge \pi. \end{cases}$$

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

$$\mathcal{L}(y'')(s) = s^2 \mathcal{L}(y) - sy(0) - y'(0) = s^2 \mathcal{L}(y) + 1,$$

check the reference   we get, making use of the previous example 7.H.10 , 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,

$$\mathcal{L}(y) = -\frac{1}{s^2 + 4} + (1 - e^{-\pi s}) \frac{s}{(s^2 + 4)^2}.$$

More details!   The inverse transform then yields the solution in the form

$$y(t) = -\tfrac{1}{2} \sin(2t) + \tfrac{1}{4} t \sin(2t) + \mathcal{L}^{-1}\left(e^{-\pi s} \frac{s}{(s^2 + 4)^2} \right).$$

According to (1),

$$\mathcal{L}^{-1}\left(\mathrm{e}^{-\pi s}\,\frac{s}{(s^2+4)^2}\right) = \tfrac{1}{4}\mathcal{L}^{-1}(\mathrm{e}^{-\pi s}\,\mathcal{L}(t\sin(2t)))$$

$$= (t-\pi)\sin(2(t-\pi))\cdot H_\pi(t).$$

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

$$y(t) = \begin{cases} \frac{t-2}{4}\sin(2t) & \text{for } 0 \le t < \pi \\ \left(\frac{5t-2}{4} - \pi\right)\sin(2t) & \text{for } t \ge \pi. \end{cases}$$

$\square$

Key to the exercises

7.A.4. We have just to check the orthogonality of the couples $L_m(x)$, $L_n(x)$ with respect to the inner product $\langle f, g \rangle_\omega = \int_0^\infty f(t)g(t) \, \mathrm{e}^{-t} \, \mathrm{d}\,t$. This can be done by integration by parts.

7.A.5. 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.6. $x, -\frac{3}{\pi^2}x + \sin(x)$; the projection does not change the function $\frac{1}{2}\sin(x)$ since it already lies in the space.

7.A.7. $\cos(x), \frac{4}{\pi}\cos(x) + x$. The projection is $3\pi/(\pi^4 - 24)(4\cos(x) + \pi x)$. Notice that this is a very bad approximation.

7.B.1. We have already checked the orthogonality of the cosine terms in the solution to the example 7.A.3. 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.B.2. Look at the previous example and use the substitution $y = \omega x$.

7.B.7. Compute exactly as in the exercise 7.B.6 and check the result against the real versions of the same Fourier series, as computed before. The complex coefficients for the real functions are always related by the complex conjugation. That is, $c_{-n} = \overline{c_n}$, see 7.1.7.

7.B.8. This is again a straightforward computation.

7.C.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.C.14. It is a good exercise on derivation and integration by parts. We may differentiate with respect to t inside the integral and $\frac{\mathrm{d}}{\mathrm{d}\,t} f(t-x)$ can be interpreted as $-\frac{\mathrm{d}}{\mathrm{d}\,x} f(t - x)$.

7.C.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.D.2. The definition of $\Gamma(t)$ reveals

$$\mathcal{L}(t^\alpha) = \int_0^\infty \mathrm{e}^{-st}\, t^\alpha \, \mathrm{d}\,t = \frac{1}{s^{\alpha+1}} \int_0^\infty \mathrm{e}^{-x}\, x^\alpha \, \mathrm{d}\,x = \frac{\Gamma(\alpha + 1)}{s^{\alpha+1}}.$$

7.D.3. Integrate by parts to obtain

$$\mathcal{L}(g)(s) = \int_0^\infty t\,\mathrm{e}^{-t}\,\mathrm{e}^{-st}\,\mathrm{d}\,t = \int_0^\infty t\,\mathrm{e}^{-(s+1)t}\,\mathrm{d}\,t = \lim_{t\to\infty}\left(\frac{t\,\mathrm{e}^{-(s+1)t}}{-(s+1)}\right) - 0 - \int_0^\infty \frac{\mathrm{e}^{-(s+1)t}}{-(s+1)}\,\mathrm{d}\,t = -\left(\lim_{t\to\infty}\frac{\mathrm{e}^{-(s+1)t}}{(s+1)^2} - \frac{\mathrm{e}^0}{(s+1)^2}\right) = \frac{1}{(s+1)^2}.$$

Differentiating the Laplace transform of a general function $-f$ (i. e., an improper integral) with respect to the parameter s gives

$$\left(\int_0^\infty -f(t)\,\mathrm{e}^{-st}\,\mathrm{d}\,t\right)' = \int_0^\infty -f(t)\left(\mathrm{e}^{-st}\right)'\,\mathrm{d}\,t = \int_0^\infty t\,f(t)\,\mathrm{e}^{-st}\,\mathrm{d}\,t.$$

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.D.4.

$$\begin{aligned} &\mathcal{L}(\cos\omega t)(s) + i\mathcal{L}(\sin\omega t)(s) = \mathcal{L}(\mathrm{e}^{i\omega t})(s) \\ =\ & \int_0^\infty \mathrm{e}^{i\omega t}\,\mathrm{e}^{-st}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{(i\omega - s)t}\,\mathrm{d}t \\ =\ & -\frac{1}{s - i\omega}\left[\mathrm{e}^{(i\omega - s)t}\right]_0^\infty \\ =\ & -\frac{1}{s - i\omega}\left(\lim_{t\to\infty}\frac{\mathrm{e}^{i\omega t}}{\mathrm{e}^{st}} - 1\right) = \frac{1}{s - i\omega} = \frac{s + i\omega}{(s - i\omega)(s + i\omega)} \\ =\ & \frac{s}{s^2 + \omega^2} + i\frac{\omega}{s^2 + \omega^2} \end{aligned}$$

7.G.5. 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.
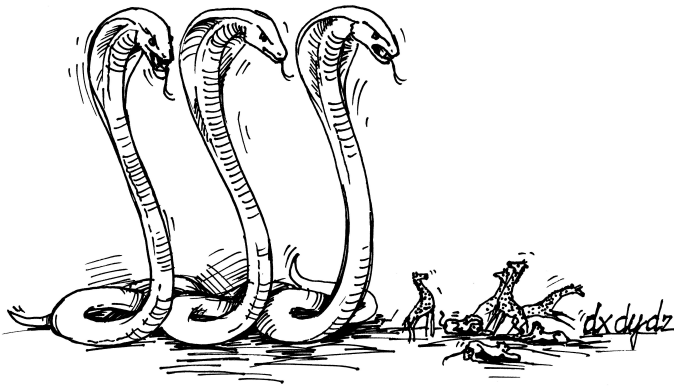
7.H.1. $\frac{1}{2} - \frac{1}{2}\cos(2x)$.

7.H.2. $\frac{1}{2} + \frac{1}{2}\cos(2x)$.

7.H.5. Determine the convolution of the functions f_1 and f_2, where

$$
f_1 * f_2(t) \quad = \quad \begin{cases} \frac{(t+1)^2}{2} & \text{for } t \in [-1, 0] \\ \frac{1-t^2}{2} & \text{for } t \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}
$$

Calculus with more variables

one variable is not enough?
– never mind, just recall vectors!

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 $\mathbb{R} \to \mathbb{R}$ 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.

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 $f(x,y) \geq 0$ ($f : \mathbb{R}^2 \to \mathbb{R}$ is a function of two variables,

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, $f : \mathbb{R}^n \to \mathbb{R}^m$. 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 $f : \mathbb{R}^2 \to \mathbb{R}^2$. We have also dealt with vector valued real functions, representing parametrized curves $c : \mathbb{R} \to \mathbb{R}^n$ (see e.g. the paragraphs on curvatures and Frenet frames in 6.1.15 on page 387).

In linear algebra and geometry, we saw the linear and affine maps $f : \mathbb{R}^n \to \mathbb{R}^m$ defined with the help of matrices $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ and constant vectors $y \in \mathbb{R}^m$:

$$\mathbb{R}^n \ni x \mapsto y + A\,x \in \mathbb{R}^m.$$

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 $\mathbb{R}^n \to \mathbb{R}$ 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 $\sum_{i,j} s_{ij}x_i x_j$.

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 solve each inequality separately and then intersect the result. In our cases we get
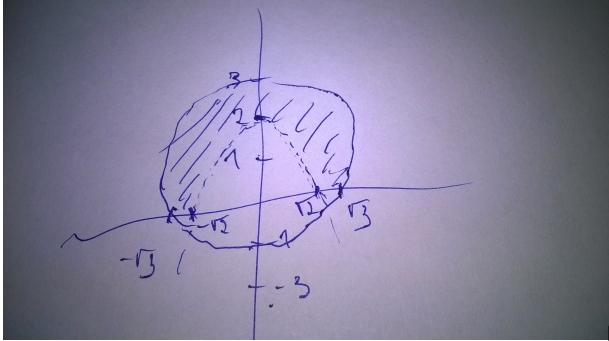
8.A.2. Determine the domain of the function $\mathbb{R}^2 \to \mathbb{R}$:

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 $\{(x, y) \in \mathbb{R}, |x| > |y|\}$. You can see the graph of this function in the picture.

In general, all such mappings $f : \mathbb{R}^n \to \mathbb{R}^m$ are composed of m components of functions $f_i : \mathbb{R}^n \to \mathbb{R}$. So we start with this case.

8.1.2. Multivariate functions. We can stress the dependance on the variables $x_1, \ldots, x_n$ by writing the functions as

$$f(x_1, x_2, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}.$$

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

$$G_f = \{(x_1, \ldots, x_n, f(x_1, \ldots, x_n)); \ (x_1, \ldots, x_n) \in A\},$$

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

$$\mathbb{R} \to \mathbb{R}^3, \ y \mapsto (x, y, f(x, y)),$$

i.e., a *curve* in the space $\mathbb{R}^3$. Curves are vector functions of one variable, already worked with in chapter six (see 6.1.13). The images of the curves for some fixed values of the coordinates x and y are depicted by lines in the illustration.
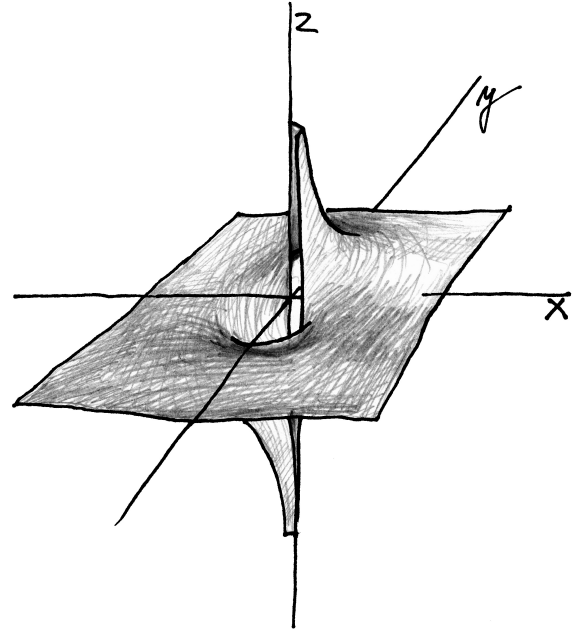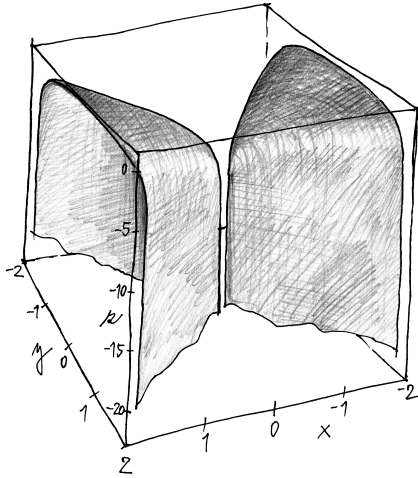
c) This formula is again a composition of a logarithm and a polynomial of two variables. However, the polynomial $-x^2 - y^2$ takes on only non-positive real values, where the logarithm is undefined (as a function $\mathbb{R} \to \mathbb{R}$).

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

$$\{[x, y, z] \in \mathbb{R}^3; (x \geq 1 \wedge y \neq 0 \wedge 0 \leq z \frac{1}{y^2})$$

$$\vee (x \in (0, 1) \wedge y \neq 0 \wedge -\frac{1}{y^2} \leq z \leq 0) \vee (x > 0 \wedge y = 0)\}.$$

$\square$

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 $(x^2 + y^2 \geq$

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 $u = (x_1, \ldots, x_n)$ and $v = (y_1, \ldots, y_n)$ 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:

$$\|P - R\| = \|(P - Q) + (Q - R)\| \leq \|P - Q\| + \|Q - R\|.$$

511

$1, x^2 + y^2 \leq 4$) 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 verticies $[\pm 1, \pm 1]$

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. $\qquad\square$

B. The topology of E_n

In the previous chapter, we have defined general metric spaces and we have studies 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 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 conecting 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

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 nad 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 δ–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 δ–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 δ–neighbourhood of P which lies inside A;

(9) *a bounded set*: lies inside some δ–neighbourhood of one of its points (for a sufficiently large δ);

(10) *a compact set*: both closed and bounded.

(11) *limit of a mapping*: $a \in \mathbb{R}^m$ is the limit of function $f : \mathbb{R}^n \to \mathbb{R}^m$ in a limit point x_0 of its domain A, if for each $\varepsilon > 0$, there is a δ–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 $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuous in $x_0 \in A$ if the limit $\lim_{x \to x_0} f(x)$ 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 compoments, 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 compoment 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

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 $E_n \setminus U$. □

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. □

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 $E_n = U \cup \overline{U}$, where both U and $\overline{U}$ (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. □

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

$$\bigcup_{i=3}^{\infty} \left[\frac{1}{i}, 1 - \frac{1}{i} \right]$$

of closed subsets of $\mathbb{R}$, which is equal to the open interval $(0, 1)$. □

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 $\{1\}$. □

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 498.

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 δ–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 δ–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. □

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

8.B.6. Consider the graph of a continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ 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.) □

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 functins, 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 easest 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 expresion, we are done. Otherwise we can get "undeterminate" expression. There are some tricks, we can use to count such a limit:

space $\mathbb{R}^n$. Several tools for the work with curves are introduced in paragraphs 6.1.13–6.1.16.

For every (parametrized) curve[1], that is, a mapping $c = (c_1(t), \dots, c_n(t)) : \mathbb{R} \to \mathbb{R}^n$ 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:*
$$\lim_{t \to t_0} c(t) = (\lim_{t \to t_0} c_1(t), \dots, \lim_{t \to t_0} c_n(t)) \in \mathbb{R}^n$$

(2) *derivative:*
$$c'(t_0) = \lim_{t \to t_0} \frac{1}{|t - t_0|} \cdot (c(t) - c(t_0))$$
$$= (c_1'(t_0), \dots, c_n'(t_0)) \in \mathbb{R}^n$$

(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

$$c_i(b) - c_i(a) = (b - a) \cdot c_i'(t_i), \quad i = 1, \dots, n.$$

---
[1] In geometry, one often makes a distinction between a curve as a subset of E_n and its parametrization $\mathbb{R} \to \mathbb{R}^n$. The word "curve", means exclusively the parametrized curve here.

(1) factorize the numerator or the denominator acording 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

$$x = r\cos\varphi,$$
$$y = r\sin\varphi$$

(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 φ);

(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 exists)

8.C.1. $\lim_{(x,y)\to(e^2,1)} \frac{\ln x}{y}$ ○

8.C.2. $\lim_{(x,y)\to(4,4)} \frac{\sqrt{x}-\sqrt{y}}{x-y}$ ○

8.C.3. $\lim_{(x,y)\to(1,\infty)} \frac{\cos y}{x+y}$ ○

8.C.4. $\lim_{(x,y)\to(0,2)} \frac{e^{xy}-1}{x}$ ○

8.C.5. $\lim_{(x,y)\to(\infty,\infty)} \frac{x^2+y^2}{x^4+y^4}$ ○

8.C.6. $\lim_{(x,y)\to(0,0)} \frac{x^2+y^2}{x+y}$ ○

8.C.7. $\lim_{(x,y)\to(0,0)} \frac{x^2-y^2}{x^2+y^2}$ ○

8.C.8. $\lim_{(x,y)\to(\infty,\infty)} \left(\frac{2xy}{x^2+y^2}\right)^{x^2}$ ○

8.C.9. $\lim_{(x,y)\to(1,1)} \frac{x+y}{\sqrt{x^2+y^2}}$ ○

8.C.10. $\lim_{(x,y)\to(0,0)} \frac{x^2+y^2}{\sqrt{x^2+y^2+1}-1}$ ○

8.C.11. $\lim_{(x,y)\to(0,0)} xy^2 \cos\frac{1}{xy^2}$ ○

8.C.12. $\lim_{(x,y)\to(0,0)} \frac{\sin xy}{x}$ ○

8.C.13. $\lim_{(x,y)\to(0,0)} \frac{x^3+y^3}{x^2+y^2}$ ○

8.C.14. $\lim_{(x,y)\to(\infty,\infty)} (x^2+y^2)e^{-(x+y)}$ ○

8.C.15. $\lim_{(x,y)\to(\infty,1)} \left(1+\frac{1}{x}\right)^{\frac{x^2}{x+y}}$ ○

8.C.16. $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$ ○

8.C.17. $\lim_{(x,y)\to(0,0)} \frac{1-\cos(x^2+y^2)}{(x^2+y^2)xy}$ ○

8.C.18. Prove that $\lim_{(x,y)\to(0,0)} \frac{-y}{x^2-y}$ does not exists. ○

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, $c(t) = (x(t), y(t))$

$$c(b) - c(a) = (x'(\xi)(b-a), y'(\eta)(b-a))$$
$$= (b-a) \cdot (x'(\xi), y'(\eta))$$

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 $a \leq s \leq t \leq b$*

$$\|c(t) - c(s)\| \leq \sqrt{n}\big(\max_{r\in[a,b]} \|c'(r)\|\big) |t - s|.$$

PROOF. Direct application of the mean value theorem gives for appropriate points r_i inside the interval $[s, t]$ the following:

$$\|c(t) - c(s)\|^2 = \sum_{i=1}^{n}(c_i(t) - c_i(s))^2 \leq \sum_{i=1}^{n}\big(c_i'(r_i)(t - s)\big)^2$$

$$\leq (t-s)^2 \sum_{i=1}^{n} \max_{r\in[s,t]} c_i'(r)^2$$

$$\leq n(\max_{r\in[s,t],\, i=1,\ldots,n} |c_i'(r)|)^2(t-s)^2$$

$$\leq n \max_{r\in[s,t]} \|c'(r)\|^2(t-s)^2. \qquad \square$$

Another important concept is the *tangent vector* to a curve $c : \mathbb{R} \to E_n$ 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

$$T: \quad c(t_0) + (t - t_0) \cdot c'(t_0)$$

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 $f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ 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 $\frac{\partial f}{\partial x_i}$, $i = 1, \ldots, n$, or (without referring to the particular function) as the operator $\frac{\partial}{\partial x_i}$ on the functions.

More generally, for every function $f : \mathbb{R}^n \to \mathbb{R}$ and an arbitrary curve $c : \mathbb{R} \to \mathbb{R}^n$, their composition $(f \circ c)(t) : \mathbb{R} \to \mathbb{R}$ can be considered. This composite function $F = f \circ c$ expresses the behaviour of the function f along the curve

8.C.19. Prove that

$$\lim_{(x,y)\to(1,-2)} \frac{2x+xy-y-2}{x^2+y^2-2x+4y+5}.$$

does not exist. ○

Solution. The lines through $[1,-2])$ have the equation $y = kx - k - 2$. As we aproach $[1,-2]$ along one of these lines, we get the limit $\frac{k}{1+k^2}$, which is different for differnt k, thus the limit does not exists. □

Let us recall, that a funtion is continuous in points, where the limit exists and is equal to to function value.

8.C.20. Find the discontinuity points of $f(x,y) = \frac{2x-5y}{x^2+y^2-1}$. ○

8.C.21. Find the discontinuity points of $f(x,y) = \frac{\sin(x^2y+xy^2)}{\cos(x-y)}$. ○

8.C.22. Find the discontinuity points of

$$f(x,y) = \begin{cases} \frac{x^3+y^3}{x^2+y^2} & \text{pro } [x,y] \neq [0,0], \\ 0 & \text{pro } [x,y] = [0,0]. \end{cases}$$

○

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

c. The simplest case is using parametrized straight lines c and choosing the lines $c_i(t) = (x_1, \ldots, x_i + t, \ldots, x_n)$, 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. $f : \mathbb{R}^n \to \mathbb{R}$ 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.

$$d_v f(x) = \lim_{t\to 0} \frac{1}{t}(f(x+tv) - f(x)).$$

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,

$$\frac{\partial}{\partial x} f(x,y) = \lim_{t\to 0} \frac{1}{t}(f(x+t,y) - f(x,y))$$

$$\frac{\partial}{\partial y} f(x,y) = \lim_{t\to 0} \frac{1}{t}(f(x,y+t) - f(x,y)).$$

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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 continous 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,

$(\cos t - 1, \sin t, t)$. 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).
□

8.D.2. Determine both the parametric and implicit equations of the tangent line to the curve $c : \mathbb{R} \to \mathbb{R}^3$, $c(t) = (c_1(t), c_2(t), c_3(t)) = (t, t^2, t^3)$ 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 the derivatives at the point $t = 1$ are 1, 2, 3. Therefore, the parametric equations of the tangent line are:

$$x = c_1'(1)s + c_1(1) = t + 1,$$
$$y = c_2'(1)s + c_2(1) = 2t + 1,$$
$$z = c_3'(1)s + c_3(1) = 3t + 1.$$

In order to get the implicit equations (which are not given canonically), we eliminate the parameter t, thereby obtaining:
$$2x - y = 1,$$
$$3x - z = 2. \qquad \square$$

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 paralell to the plane $\varrho\colon 3x + y - z + 7 = 0$.

Solution. The direction $c'(t_0)$ of the curve $c(t)$ in t_0 has to be perpendicular to the normal of ϱ, 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 ρ is $(3, 1, -1)$ (just read off the coefficients by x, y and z in the equation of ρ. That is $3 \cdot 2t + 1 \cdot (-4t + 5) + 1 \cdot 1 = 0$, which gives $[3, -18, -7]$. $\qquad \square$

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. $p = \{[1 - \sqrt{2}s, 1, \sqrt{2} + s]; s \in \mathbb{R}\}$. $\qquad \square$

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 θ, takes the values $\alpha(\theta)r$, where $\alpha(\theta)$ is a periodic odd function of the angle θ, with period 2π. 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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) $\lim_{v \to 0} \frac{1}{\|v\|} \left(f(x + v) - f(x) - d_v f(x) \right) = 0$.

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.

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 $x_0 = x(0)$, $y_0 = y(0)$.

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.
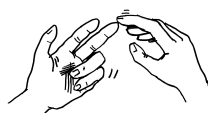
8.D.6. The set of differentiable functions. We can notice that multivariate polynomials are differentiable on the whole of their domain. Similarly, the composition of a differentiable univariate function and a differentiable multivariate function leads to a differentiable multivariate function. For instance, 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 $f : \mathbb{R}^n \to \mathbb{R}$ in variables $x_1, \ldots, x_n$ 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 $f : \mathbb{R}^2 \to R$, $f(x, y) = x^2\sqrt{y}$. Calculate the partial derivatives where they are defined on this domain.

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 $y = 0$. □

8.D.8. Determine the derivative of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2yz$ 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 $\mathbb{R}^3$.

Apply this in both summands of the following expression separately, to obtain

$$\tfrac{1}{t}\big(f(x(t), y(t)) - f(x_0, y_0)\big) =$$

$$\tfrac{1}{t}\big(f(x(t), y(t)) - f(x_0, y(t))\big) + \tfrac{1}{t}\big(f(x_0, y(t)) - f(x_0, y_0)\big)$$

$$= \tfrac{1}{t}(x(t) - x_0) \cdot \frac{\partial f}{\partial x}(x(\xi), y(t)) + \tfrac{1}{t}(y(t) - y_0) \cdot \frac{\partial f}{\partial y}(x_0, y(\eta))$$

for suitable numbers ξ and η between 0 and t. Indeed, by exploiting that the curve $(x(t), y(t))$ is differentiable, there must be such values ξ and η.

Especially, for every sequence of numbers t_n converging to zero, the corresponding sequences of numbers ξ_n and η_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 $\mathbb{R}^n \to \mathbb{R}$ 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.

8.1.9. The following theorem provides a crucial and very useful observation.

Let us follow the definition:

$$f_v(x, y, z) = \lim_{t \to 0} \frac{1}{t}[f(x + 3t, y + 2t, z - t) - f(x, y, z)]$$
$$= \lim_{t \to 0} \frac{1}{t}[(x + 3t)^2(y + 2t)(z - t) - x^2yz]$$
$$= \lim_{t \to 0} \frac{1}{t}[t(6xyz + 2x^2z - x^2y) + t^2(\dots)]$$
$$= 6xyz + 2x^2z - x^2y.$$

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:

$$f_x = 2xyz, \ f_y = x^2z, \ f_z = x^2y.$$

It follows from the note beyond theorem 8.1.8 that we can express

$$f_v(1, -1, 2) = 3f_x(1, -1, 2) + 2f_y(1, -1, 2) +$$
$$+ (-1)f_z(1, -1, 2) =$$
$$= 3 \cdot (-4) + 2 \cdot 2 + (-1) \cdot (-1) = -7. \qquad \square$$

8.D.9. Determine the derivative of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = \frac{\cos(x^2y)}{z}$ 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 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

$$f_x(0, 0, 2) + 2 \cdot f_y(0, 0, 2) + 3 \cdot f_z(0, 0, 2)$$

$$= 1 \cdot 0 + 2 \cdot 0 + 3 \cdot \left(-\frac{1}{4}\right) = -\frac{3}{4}. \qquad \square$$

Theorem. *Let $f : E_n \to \mathbb{R}$ 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:

$$f(x_1 + c_1(t), \dots, x_n + c_n(t)) - f(x_1, x_2 + c_2(t), \dots)$$
$$+ f(x_1, x_2 + c_2(t), \dots)) - f(x_1, x_2, \dots, x_n + c_n(t))$$
$$\vdots$$
$$+ f(x_1, x_2, \dots, x_n + c_n(t)) - f(x_1, x_2, \dots, x_n).$$

Now, apply the mean value theorem to all of the n summands, obtaining (similarly to the case of two variables)

$$(c_1(t) - c_1(0))\frac{\partial f}{\partial x_1}(x_1 + c_1(\theta_1), \dots, x_n + c_n(t))$$
$$+ (c_2(t) - c_2(0))\frac{\partial f}{\partial x_2}(x_1, x_2 + c_2(\theta_2), \dots, x_n + c_n(t))$$
$$\vdots$$
$$+ (c_n(t) - c_n(0))\frac{\partial f}{\partial x_n}(x_1, x_2, \dots, x_n + c_n(\theta_n)),$$

for appropriate values θ_i, $0 \le \theta_i \le t$. 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

$$f(x + v) - f(x) = d_vf(x + \theta v)$$
$$= v_1\frac{\partial f}{\partial x_1}(x + \theta v) + \dots + v_n\frac{\partial f}{\partial x_n}(x + \theta v)$$

with an appropriate θ, $0 \le \theta \le 1$, 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

8.D.10. Having a function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f_v(x) = v_1 f_{x_1}(x) + v_2 f_{x_2}(x) + \cdots + v_n f_{x_n}(x)$ in dependence on the variables $v_1, \ldots, v_n$ which are bound by the condition $v_1^2 + \cdots + v_n^2 = 1$. We have already solved this type of problem in chapter 3, when we talked about linear optimization (viz 3.A.1). The value $f_v(x)$ can be interpreted as the scalar product of the vectors $(f_{x_1}, \ldots, f_{x_n})$ and $(v_1, \ldots, v_n)$. And this product is maximal if the vectors are of the same direction. The vector v can thus be obtained by normalizing the vector $(f_{x_1}, \ldots, f_{x_n})$. In general, we say the the function grows maximally in the direction $(f_{x_1}, \ldots, f_{x_n})$. Then, this vector is called the gradient of the function f. In paragraph 8.1.25, we will recall this idea and go into further details. $\square$

Counnting 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.

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 $f'_x = \frac{x}{\sqrt{x^2+y^2}}$,

δ–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 ε. Hence the estimate

$$\frac{1}{\|w\|}\big\|f(x + w) - f(x) - d_w f(x)\big\| \leq$$

$$\leq \frac{1}{\|w\|}\big\|f(x + w) - f(x) - d_w f(x + \theta w)\big\| +$$

$$\frac{1}{\|w\|}\big\|d_w f(x + \theta w) - d_w f(x)\big\|$$

$$= \frac{1}{\|w\|}\Big\|w_1\big(\frac{\partial f}{\partial x_1}(x + \theta w) - \frac{\partial f}{\partial x_1}(x)\big) + \ldots$$

$$+ w_n\big(\frac{\partial f}{\partial x_n}(x + \theta w) - \frac{\partial f}{\partial x_n}(x)\big)\Big\|$$

$$\leq \frac{n}{\|w\|}\|w\|\varepsilon,$$

where θ 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 $\lim_{v \to 0} \frac{\alpha(v)}{\|v\|} = 0$, i.e. $\alpha(v) = o(\|v\|)$ in the asymptotic terminology introduced in 6.1.16 on the page 389.

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):

$$z = f(x_0, y_0) + df(x_0, y_0)(x - x_0, y - y_0)$$

$$= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0).$$

It is already seen that the increase of the function values of a differentiable function $f : E_n \to \mathbb{R}$ 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.

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 $c(t) = (t, t, f(t, t))$.

$f'_y = \frac{y}{\sqrt{x^2+y^2}}$, thus

$$\sqrt{2.98^2 + 4.05^2} \doteq f(3^2, 4^2) + df(2.98 - 3, 4.05 - 4)$$

$$= \sqrt{3^2 + 4^2} + \frac{3}{\sqrt{3^2 + 4^2}}(-0.02) + \frac{4}{\sqrt{3^2 + 4^2}}(0.05)$$

$$= 5 - \frac{0,06}{5} + \frac{0,2}{5} = 5,028.$$

$\square$

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) $1,04^{2,02}$.

$\bigcirc$

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 fuction 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 diffefential 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 $f : \mathbb{R}^2 \to \mathbb{R}$ in a point $P = [x_0, y_0, f(x_0, y_0)$:
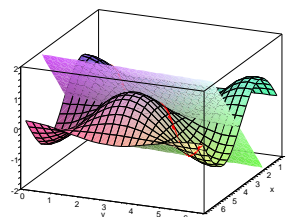
i) $f(x, y) = \sqrt{1 - x^2 - y^2}$, $P = [x_0, y_0, z_0] = [\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, ?]$.

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) $z_0 = -\frac{\pi}{4}, x + y - 2z = \frac{\pi}{2}$.
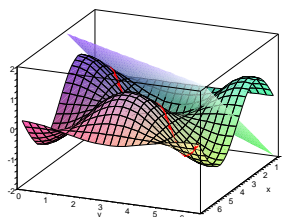
For the case of functions of n variables, the tangent plane is defined as an analogy to the tangent plane to a surface in the three-dimensional space. Instead of being overwhelmed by many indices, it is useful to recall affine geometry, where hyperplanes can be used, see paragraph 4.1.3.

TANGENT (HYPER)PLANES

Definition. A *tangent hyperplane* to the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $df(x) : \mathbb{R}^n \to \mathbb{R}$, 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 $f : E_n \to \mathbb{R}$

$$f \mapsto d_v f = df(v),$$

and the result is again a function $df(v) : E_n \to \mathbb{R}$. If this function is differentiable as well, repeat this procedure with another increment, and so on. In particular, work with iterations of partial derivatives. For *second-order partial derivatives*, write

$$\left(\frac{\partial}{\partial x_j} \circ \frac{\partial}{\partial x_i}\right) f = \frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

□

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 axe. 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 axe iff the tangent line in the point is parallel to the x axe. 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 eqution 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 saught points are $[1, 1]$, resp. $[1, -3]$, the equations of tangents in the points are $y = 1$, resp. $y = -3$. □

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. ○

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. ○

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 $x + y - z - u = 0$. ○

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

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*

$$\frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}}.$$

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 $f : E_n \to \mathbb{R}$ 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, supposing the second-order partial derivatives are continuous. In the plane as well as in the space, iterated derivatives are often denoted by mere indices referring to the variable names, for example:

$$f_x = \frac{\partial f}{\partial x}, \ f_{xx} = \frac{\partial^2 f}{\partial x^2}, \ f_{xy} = \frac{\partial^2 f}{\partial x \partial y}, \ f_{yx} = \frac{\partial^2 f}{\partial y \partial x}.$$

We show that the continuous partial derivatives commute. That is, the order in which differentiation is carried out does not matter.

Suppose that the partial derivatives exist and are continuous, i.e., the limits

$$f_{xy}(x, y) = \lim_{t \to 0} \frac{1}{t}\left(f_x(x, y + t) - f_x(x, y)\right)$$

$$= \lim_{t \to 0} \frac{1}{t}\left(\lim_{s \to 0} \frac{1}{s}\left(f(x + s, y + t) - f(x, y + t)\right.\right.$$

$$\left.\left. - (f(x + s, y) - f(x, y))\right)\right)$$

exist. However, since the limits can be expressed by any choice of values $t_n \to 0$ and $s_n \to 0$ and the limits of the corresponding sequences, the second derivative can be expressed as

$$f_{xy}(x, y) = \lim_{t \to 0} \frac{1}{t^2}\left(\left(f(x + t, y + t) - f(x, y + t)\right)\right.$$

$$\left. - \left(f(x + t, y) - f(x, y)\right)\right),$$

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 $g(x, y) =$

given implicitly by $F(x, y, z) = 0$ is the vector $(F'_x(x_0, y_0, z_0), F'_y(x_0, y_0, z_0), F'_z(x_0, y_0, z_0))$. $\qquad\square$

8.D.22. Determine whether the tangent plane to the graph of the function $f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$, $f(x, y) = x \cdot \ln(y)$ 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

$$z = f\left(1, \frac{1}{e}\right) + f_x\left(1, \frac{1}{e}\right)(x - 1) + f_y\left(1, \frac{1}{e}\right)\left(y - \frac{1}{e}\right)$$
$$= -1 - x + ey.$$

The given point does not satisfy this equation, so it does not lie in the tangent plane. $\qquad\square$

8.D.23. Determine the parametric equation of the tangent line to the intersection of the graphs of the functions $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + xy - 6$; $g : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$, $g(x, y) = x \cdot \ln(y)$ 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 $[2, 1 + t, 2t]$, $t \in \mathbb{R}$. $\qquad\square$

8.D.24. V. ypočtěte všechny parciální derivace prvního a druhého řádu funkce $f(x, y, z) = x^{\frac{y}{z}}$.

Solution. $f'_x = \frac{y}{z} x^{\frac{y}{z} - 1}$, $f'_y = x^{\frac{y}{z}} \ln x \cdot \frac{1}{z}$, $f'_z = x^{\frac{y}{z}} \ln x \cdot \frac{-y}{z^2}$, $f''_{xx} = \frac{y}{z}(\frac{y}{z} - 1)x^{\frac{y}{z} - 2}$, $f''_{yy} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{1}{z^2}$,

$f(x + t, y) - f(x, y)$. Then the expression in the last large parentheses is, by the mean value theorem, equal 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).

Now, $g_y(x, y) = f_y(x + t, y) - f_y(x, y)$, so we may rewrite φ as

$$\varphi(x, y, t) = \frac{1}{t} g_y(x, y + t_0)$$
$$= \frac{1}{t}\big(f_y(x + t, y + t_0) - f_y(x, y + t_0)\big).$$

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. However, if the large parentheses are split into $(f(x + t, y + t) - f(x + t, y)) - (f(x, y + t) - f(x, y))$, we get, by the same procedure with the function $h(x, y) = f(x, y + t) - f(x, y)$, the expression

$$\varphi(x, y, t) = f_{xy}(x + s_0, y + s_1)$$

with (in general) different constants s_0 and s_1. Since it is assumed that the second-order partial derivatives are continuous, the limit for $t \to 0$ guarantees the desired equality

$$f_{xy}(x, y) = f_{yx}(x, y)$$

at all points (x, y).

8.1.12. The same procedure for functions of n variables proves the following fundamental result:

COMMUTATIVITY OF PARTIAL DERIVATIVES

Theorem. *Let $f : E_n \to \mathbb{R}$ be a k–times differentiable function with continuous partial derivatives up to order k (inclusive) in a neighbourhood of a 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 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 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 $i_1, \ldots, i_k$ can be obtained from a fixed one by several interchanges of adjacent pairs of indices. $\qquad\square$

$f''_{zz} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{y^2}{z^4} + x^{\frac{y}{z}} \ln x \cdot \frac{2y}{z^3}$,

$f''_{xy} = \frac{1}{z} x^{\frac{y}{z}-1} + \frac{y}{z} x^{\frac{y}{z}-1} \ln x \cdot \frac{1}{z}, f''_{xz} = \frac{-y}{z^2} x^{\frac{y}{z}-1} + \frac{y}{z} x^{\frac{y}{z}-1} \ln x \cdot \frac{-y}{z^2}, f''_{yz} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{-y}{z^3} + x^{\frac{y}{z}} \ln x \cdot \frac{-1}{z^2}$.

□

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 $x^2 + y^2 + z^2 - xz - \sqrt{2}yz = 1$.

○

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 $2x^2 + 2y^2 + z^2 + 8xz - z + 8 = 0$.   ○

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, $Df = \{(x, y, z) \in \mathbb{R}^3, (z \geq 1 \& (xy > 0)) \vee (0 < z < 1) \& (xy < 0)\}$.

Now, we calculate the first partial derivatives with respect to each of the three variables:

$f_x = \dfrac{y \ln(z)}{2\sqrt{xy \ln(z)}}, f_y = \dfrac{x \ln(z)}{2\sqrt{xy \ln(z)}}, f_z = \dfrac{xy}{2z\sqrt{xy \ln(z)}}$.

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.

$f_{xx} = -\dfrac{y^2 \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}}$,

$f_{xy} = -\dfrac{xy \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}} + \dfrac{\ln z}{2\sqrt{xy \ln z}}$,

$f_{xz} = -\dfrac{xy^2 \ln z}{4z(xy \ln z)^{\frac{3}{2}}} + \dfrac{y}{2z\sqrt{xy \ln z}}$,

$f_{yy} = -\dfrac{x^2 \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}}$,

$f_{yz} = -\dfrac{x^2 y \ln z}{4z(xy \ln z)^{\frac{3}{2}}} + \dfrac{x}{2z\sqrt{xy \ln z}}$,

$f_{zz} = -\dfrac{x^2 y^2}{4z^2(xy \ln z)^{\frac{3}{2}}} - \dfrac{xy}{2z^2\sqrt{xy \ln z}}$.

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 $f : E_n \to \mathbb{R}$ is possible.

HESSIAN

Definition. If $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function, the symmetric matrix of functions

$$Hf(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}$$

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

$$\beta(t) = f(x_0, y_0) + t\frac{\partial f}{\partial x}(x_0, y_0)\xi + t\frac{\partial f}{\partial y}(x_0, y_0)\eta$$

$$+ \frac{t^2}{2}\left(f_{xx}(x_0, y_0)\xi^2 + 2f_{xy}(x_0, y_0)\xi\eta + f_{yy}(x_0, y_0)\eta^2 \right)$$

shares the same derivatives up to the second order (inclusive) at the point $t = 0$ (calculate this on your own!). The function β 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 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 obrazku

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.
□

E. Taylor polynomials

8.E.1. Write the second-order Taylor expansion of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \ln(x^2 + y^2 + 1)$ 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

$$\begin{pmatrix} \frac{2}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{2}{9} \end{pmatrix}.$$

Altogether, we get that the second-order Taylor expansion at the point $[1, 1]$ is

$$\begin{aligned} T_2(x, y) = & f(1, 1) + f_x(1, 1)(x - 1) + f_y(1, 1)(y - 1) \\ & + \frac{1}{2}(x - 1, y - 1) Hf(1, 1) \begin{pmatrix} x - 1 \\ y - 1 \end{pmatrix} \\ = & \ln(3) + \frac{2}{3}(x - 1) + \frac{2}{3}(y - 1) + \frac{1}{9}(x - 1)^2 \\ & - \frac{4}{9}(x - 1)(y - 1) + \frac{1}{9}(y - 1)^2 \\ = & \frac{1}{9}(x^2 + y^2 + 8x + 8y - 4xy - 14) + \ln(3). \end{aligned}$$

□

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 $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x, y) = xy \cos y$ 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

$$T(x, y) = \frac{1}{2}\pi^2 y^2 - xy - \pi^3 y + \frac{1}{2}\pi^4.$$
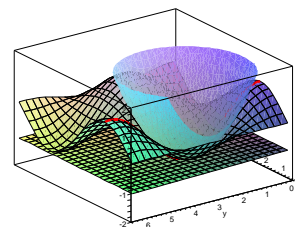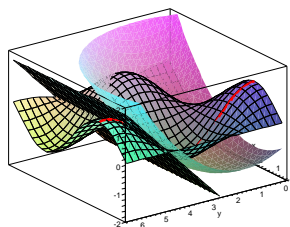
8.1.14. Taylor's expansion. The multidimensional version of Taylor's theorem is an example of a mathematical statement where the most difficult part is finding the right formulation. The proof is then quite simple.

The discussion on the Hessians continues. Write $D^k f$ for the k–th order approximations of the function $f : E_n \to \mathbb{R}^n$. 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 $f : E_n \to \mathbb{R}$, points $x = (x_1, \ldots, x_n) \in E_n$, and increments $v = (\xi_1, \ldots, \xi_n)$, set

$$D^k f(x)(v) = \sum_{1 \le i_1, \ldots, i_k \le n} \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x_1, \ldots, x_n) \xi_{i_1} \cdots \xi_{i_k}.$$

An illustrative example (making use of the symmetry of the partial derivatives) is, for E_2, the third-order expression

$$\begin{aligned} D^3 f(x, y)(\xi, \eta) = & \frac{\partial^3 f}{\partial x^3} \xi^3 + 3 \frac{\partial^3 f}{\partial x^2 \partial y} \xi^2 \eta \\ & + 3 \frac{\partial^3 f}{\partial x \partial y^2} \xi \eta^2 + \frac{\partial^3 f}{\partial y^3} \eta^3, \end{aligned}$$

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 EXPANSION WITH REMAINDER

Theorem. *Let $f : E_n \to \mathbb{R}$ 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 θ, $0 \le \theta \le 1$, such that*

$$f(x + v) = f(x) + D^1 f(x)(v) + \frac{1}{2!} D^2 f(x)(v) +$$

$$\cdots + \frac{1}{(k - 1)!} D^{k-1} f(x)(v) + \frac{1}{k!} D^k f(x + \theta \cdot v)(v).$$

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 $\varphi : \mathbb{R} \to \mathbb{R}$ defined by the composition $\varphi(t) = f \circ c(t)$. Taylor's theorem for univariate functions claims that (see Theorem 6.1.3)

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 $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^3 y + xz^2 + xy + 1$ at the point $[0, 0, 0]$. ◯

8.E.4. Determine the second-order Taylor polynomial of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 \sin y + y^2 \cos x$ 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 $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x, y) = \tan(xy + y)$$

at the point $[0, 0]$. ◯

F. Extrema of multivariate functions

8.F.1. Determine the stationary points of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 y + y^2 x - xy$ 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 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,
$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

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. □

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \dots$$
$$+ \frac{1}{(k-1)!} \varphi^{(k-1)}(0) t^{k-1} + \frac{1}{k!} \varphi^{(k)}(\theta) t^k.$$

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

$$\varphi^{(\ell)}(t) = D^\ell f(x(t))(v)$$
$$= \sum_{i_1, \dots, i_\ell} \left(\frac{\partial^\ell f}{\partial x_{i_1} \dots \partial x_{i_\ell}}(x_1(t), \dots, x_n(t)) x_{i_1}'(t) \cdots x_{i_\ell}'(t) \right)$$

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

$$\varphi^{(\ell+1)}(t) = \frac{d}{dt} D^\ell f(c(t))(c'(t))$$
$$= \frac{d}{dt} \sum_{i_1, \dots, i_\ell} \left(\frac{\partial^\ell f}{\partial x_{i_1} \dots \partial x_{i_\ell}}(x_1(t), \dots, x_n(t)) \right.$$
$$\left. \cdot x_{i_1}'(t) \cdots x_{i_\ell}'(t) \right)$$
$$= \sum_{i_1, \dots, i_\ell} \left(\sum_{j=1}^n \frac{\partial^{\ell+1} f}{\partial x_{i_1} \dots \partial x_{i_\ell} \partial x_j}(x_1(t), \dots, x_n(t)) \right.$$
$$\left. \cdot x_j'(t) \cdot x_{i_1}'(t) \cdots x_{i_\ell}'(t) \right) + 0,$$

which is the required formula for order $\ell + 1$. Taylor's theorem now follows from the enumeration at the point $t = 0$ and substituting into the equality for φ at the beginning of this proof. □

8.1.15. Formula with multi-indices. To simplify the notation, let us introduce the multi-index notation for the polynomials with more variables.

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. □

8.F.3. Determine the local extrema of the function

$$f(x,y) = x^2 + \arctan^2 x + \left| y^3 + y \right|, \quad x,y \in \mathbb{R}.$$

Solution. The function f can be written as the sum $f_1 + f_2$, where $f_1(x) = x^2 + \arctan^2 x$, $f_2(y) = \left| y^3 + y \right|$, $x,y \in \mathbb{R}$.

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 have thus proved that f can have a local extremum only at the origin. Since

$$f(0,0) = 0, \qquad f(x,y) > 0, \quad [x,y] \in \mathbb{R}^2 \smallsetminus \{[0,0]\},$$

A *multi-index* α of length n is an n-tuple of non-negative integers $(\alpha_1, \ldots, \alpha_n)$. The integer $|\alpha| = \alpha_1 + \cdots + \alpha_n$ is called the *size* of the multi-index α.

Monomials are written shortly as x^α instead of $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$. Real polynomials in n variables can be symbolically expressed in a similar way as univariate polynomials:

$$f = \sum_{|\alpha| \leq k} a_\alpha x^\alpha, \ g = \sum_{|\beta| \leq \ell} b_\beta x^\beta \in \mathbb{R}[x_1, \ldots, x_n].$$

f is said to have total degree k if at least one coefficient with multi-indices α 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 ℓ respectively:

$$f + g = \sum_{|\alpha| \leq \max(k,\ell)} (a_\alpha + b_\alpha) x^\alpha,$$

$$fg = \sum_{|\gamma|=0}^{k+\ell} \left(\sum_{\alpha+\beta=\gamma} (a_\alpha b_\beta) x^\gamma \right),$$

where the multi-indices are added componentwise, and the formally non-existing coefficients are assumed to be zero.

Moreover we write shortly

$$\partial_\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}$$

and $\alpha! = \alpha_1! \cdots \alpha_n!$. In particular, $\partial_\alpha f = f$ if $\alpha = 0$.

TAYLOR POLYNOMIALS VIA MULTI-INDICES

Taylor expansion up to order r of a function $f : \mathbb{R}^n \to \mathbb{R}$, for an increment $v \in \mathbb{R}^n$ is the polynomial

$$(1) \qquad f(x+v) = f(x) + \sum_{1 \leq |\alpha| \leq k} \frac{1}{\alpha!} \partial_\alpha f(x) v^\alpha,$$

quite as the formula in dimension one.

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 α. Indeed, we may estimate $\sum_{|\alpha|=k} \frac{1}{\alpha!} \partial_\alpha f(x) v^\alpha \leq \frac{1}{k!} C \|nv\|^{|\alpha|}$ and thus the Taylor series converges by the Weierstrass criterion. Think about the details. Actually, our argument shows that the Taylor series converges (on a small neighborhood of x) if the partial derivatives do not grow with the order k faster than $k!$.

the function f has a strict local (even global) minimum at the point $[0,0]$. □

8.F.4. Examine the local extrema of the function

$$f(x,y) = \left(x + y^2\right) e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

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

$$f_x(x,y) = e^{\frac{x}{2}} + \frac{1}{2}\left(x + y^2\right) e^{\frac{x}{2}},$$

$$f_y(x,y) = 2y\, e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

A stationary point $[x,y]$ must satisfy

$$f_y(x,y) = 0, \quad \text{i. e.} \quad y = 0,$$

and, further,

$$f_x(x,y) = f_x(x,0) = e^{\frac{x}{2}}\left(1 + \tfrac{1}{2}x\right) = 0, \quad \text{i. e.} \quad x = -2.$$

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

$$f_{xx}(x,y) = \tfrac{1}{2} e^{\frac{x}{2}}\left(2 + \tfrac{1}{2}\left(x + y^2\right)\right), \quad f_{yy}(x,y) = 2\, e^{\frac{x}{2}},$$

$$f_{xy}(x,y) = f_{yx}(x,y) = y\, e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

Therefore,

$$Hf\left(-2,0\right) = \begin{pmatrix} f_{xx}\left(-2,0\right) & f_{xy}\left(-2,0\right) \\ f_{yx}\left(-2,0\right) & f_{yy}\left(-2,0\right) \end{pmatrix}$$

$$= \begin{pmatrix} 1/2e & 0 \\ 0 & 2/e \end{pmatrix}.$$

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 $[-2,0]$. □

8.F.5. Find the local extrema of the function

$$f(x,y,z) = x^3 + y^2 + \tfrac{z^2}{2} - 3xz - 2y + 2z, \quad x, y, z \in \mathbb{R}.$$

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

8.1.16. 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).
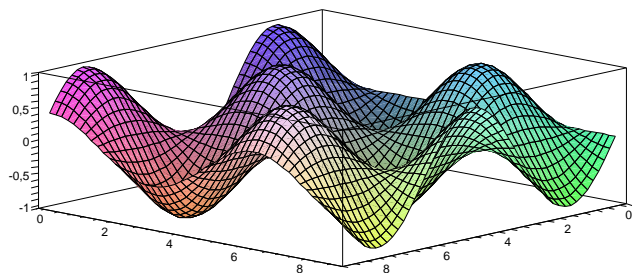
Calculate the first derivatives, and then the necessary second-order ones:

$$f_x(x,y) = \cos(x)\cos(y), \; f_y(x,y) = -\sin(x)\sin(y),$$

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:

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

$$3x^2 - 3z = 0, \quad \text{i. e.,} \quad z = x^2,$$
$$2y - 2 = 0, \quad \text{i. e.,} \quad y = 1,$$

and (utilizing the first equation)

$$z - 3x + 2 = 0, \quad \text{i. e.,} \quad x \in \{1, 2\}.$$

Therefore, there are two stationary points, namely $[1, 1, 1]$ and $[2, 1, 4]$. Now, we compute all second-order partial derivatives:

$$f_{xx} = 6x, \quad f_{xy} = f_{yx} = 0, \quad f_{xz} = f_{zx} = -3,$$
$$f_{yy} = 2, \quad f_{yz} = f_{zy} = 0, \quad f_{zz} = 1.$$

Having this, we are able to evaluate the Hessian at the stationary points:

$$Hf(1, 1, 1) = \begin{pmatrix} 6 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix},$$

$$Hf(2, 1, 4) = \begin{pmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix}.$$

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

$$d_1 = |a_{11}|, \, d_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \, d_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix}, \, \ldots$$

$$d_n = |A|,$$

are positive. Further, it is negative definite iff

(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 ℓ 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 ℓ 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 continous), 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 ℓ 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.

8.1.17. 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 4.2.6–4.2.7 in the chapter on affine geometry. There are introduced the following types of quadratic forms $h : E_n \to \mathbb{R}$:

- *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 $f(v) < 0$ for appropriate $u, v \in V$.

$$d_1 < 0, \quad d_2 > 0, \quad d_3 < 0, \quad \ldots, \quad (-1)^n d_n > 0.$$

The inequalities

$$\big|12\big| = 12 > 0, \quad \begin{vmatrix} 12 & 0 \\ 0 & 2 \end{vmatrix} = 24 > 0, \quad \begin{vmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{vmatrix} = 6 > 0,$$

imply that the matrix $Hf\,(2,1,4)$ is positive definite – there is a strict local minimum at the point $[2,1,4]$. $\qquad\square$

8.F.6. Find the local extrema of the function

$$z = \left(x^2 - 1\right)\left(1 - x^4 - y^2\right), \quad x, y \in \mathbb{R}.$$

Solution. Once again, we calculate the partial derivatives z_x, z_y and set them equal to zero. This leads to the equations

$$-6x^5 + 4x^3 + 2x - 2xy^2 = 0, \quad \left(x^2 - 1\right)(-2y) = 0,$$

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

$$z_{xx} = -30x^4 + 12x^2 + 2 - 2y^2,$$

$$z_{xy} = z_{yx} = -4xy,$$

$$z_{yy} = -2\left(x^2 - 1\right)$$

and evaluate the Hessian at the stationary points:

$$Hz\,(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

$$Hz\,(1,0) = Hz\,(-1,0) = \begin{pmatrix} -16 & 0 \\ 0 & 0 \end{pmatrix}.$$

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 in 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 $y = \sqrt{2\left(1 - x^4\right)}$, so that $y \to 0$ for $x \to \pm 1$. For this choice, we get $z\left(x, \sqrt{2\left(1 - x^4\right)}\right) = \left(x^2 - 1\right)\left(x^4 - 1\right) > 0$, $x \in (-1,1)$. 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. $\qquad\square$

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:

Theorem. *Let $f : E_n \to \mathbb{R}$ 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 minimum 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 $f(x_1, \ldots, x_n)$, an arbitrary point $x = (x_1, \ldots, x_n)$, and any increment $v = (v_1, \ldots, v_n)$, 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 θ, $0 \le \theta \le 1$. 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 4.2.7. 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. $\qquad\square$

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

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

$$p(x,y) = x^6\left(1 - y^5\right) + y^8 + y^4x^4 > 0$$

for $[x,y] \in \mathbb{R} \times (-1,1) \smallsetminus \{[0,0]\}$. Therefore, the given polynomial has a local minimum at the origin. $\square$

8.F.8. Determine local extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = x^2y + y^2z + x - z$ on $\mathbb{R}^3$. $\bigcirc$

8.F.9.

Determine the local extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = x^2y - y^2z + 4x + z$ on $\mathbb{R}^3$. $\bigcirc$

8.F.10. Determine the local extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = xz^2 + y^2z - x + y$ on $\mathbb{R}^3$. $\bigcirc$

8.F.11. Determine the local extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = y^2z - xz^2 + x + 4y$ on $\mathbb{R}^3$. $\bigcirc$

8.F.12. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2y + x^2 + 2y^2 + y$ on $\mathbb{R}^2$ $\bigcirc$

8.F.13. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2y + 2y^2 + 2y$ on $\mathbb{R}^2$. $\bigcirc$

8.F.14. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2 + xy + 2y^2 + y$ on $\mathbb{R}^2$. $\bigcirc$

8.F.15. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2 + xy - 2y^2 + y$ on $\mathbb{R}^2$. $\bigcirc$

G. Implicitly given functions and mappings

8.G.1. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be a function, $F(x,y) = xy\sin\left(\frac{\pi}{2}xy^2\right)$. Show that the equality $F(x,y) = 1$ implicitly defines a function $f : U \to \mathbb{R}$ 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 $[1,1]$:

$$F_y(x,y) = x\sin\left(\frac{\pi}{2}xy^2\right) + \pi x^2y^2\cos\left(\frac{\pi}{2}xy^2\right),$$

similar consequences as the non-vanishing of the second derivative of a univariate function. For a function $f : \mathbb{R}^n \to \mathbb{R}$, 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.

8.1.18. The differential of mappings. The concepts of derivative and differential can be easily extended to mappings $F : E_n \to E_m$. Having selected the Cartesian coordinate system on both sides, this mapping is an ordinary m–tuple

$$F(x_1,\ldots,x_n) = (f_1(x_1,\ldots,x_n),\ldots,f_m(x_1,\ldots,x_n))$$

of functions $f_i : E_n \to \mathbb{R}$. 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 $f_1,\ldots,f_m$.

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 $D^1F(x) : \mathbb{R}^n \to \mathbb{R}^m$ between the modelling spaces which linearly approximates the increments of the mapping F.

DIFFERENTIAL AND JACOBI MATRIX

Consider a differentiable mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ with components $(f_1(x_1,\ldots,x_n),\ldots,f_m(x_1,\ldots,x_n))$ 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 $v = (v_1,\ldots,v_n)$ by the Jacobi matrix is called the *differential of the mapping F* at a point x in the domain if and only if

$$\lim_{v \to 0} \frac{1}{\|v\|}\left(F(x+v) - F(x) - D^1F(x)(v)\right) = 0.$$

Recall that the definition of Euclidean distance guarantees that the limits of values in E_n exist if and only if the

so $F_y(1,1) = 1 \neq 0$. Therefore, it follows from theorem 8.1.24 that the equation $F(x,y) = 1$ implicitly determines on a neighborhood of the point $(1,1)$ a function $f : U \to \mathbb{R}$ 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

$$f'(1) = -\frac{F_x(1,1)}{F_y(1,1)} = -\frac{1}{1} = -1. \qquad \square$$

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 $F : \mathbb{R}^2 \to \mathbb{R}$,

$$F(x,y) = e^x \sin(y) + y - \pi/2 - 1,$$

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 $f : U \to \mathbb{R}$ 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 0 satisfies:

$$f'(0) = -\frac{F_x(0, \pi/2)}{F_y(0, \pi/2)} = -\frac{1}{1} = -1. \qquad \square$$

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 $z(x,y) : \mathbb{R}^2 \to \mathbb{R}$ 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 $[\pi, 1, 0]$.

$$F_x(x,y,z) = y\cos(xy) + z\cos(xz) \quad F_x(\pi,1,0) = -1,$$
$$F_y(x,y,z) = x\cos(xy) + z\cos(yz) \quad F_y(\pi,1,0) = -\pi,$$

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 $F : E_n \to E_m$ 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 $D^1 F(x)$ exists, and it is given by the Jacobi matrix $D^1 F(x)$.*

8.1.19. 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 theorem with remainder in order one on each of the components of $F = (f_1, \ldots, f_n)$ separately gives the estimate (write $v = y - x$)

$$\|F(y) - F(x)\|^2 = \sum_{i=1}^{m} |f_i(y) - f_i(x)|^2$$
$$= \sum_{i=1}^{m} |D^1 f_i(x + \theta_i v)(v)|^2$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} \left|\frac{\partial f_i}{\partial x_j}(x + \theta_i v)v_j\right|^2$$
$$\leq \left(\max_{z \in K, i, j} \left|\frac{\partial f_i}{\partial x_j}(z)\right|^2\right) nm\|v\|^2 = C^2\|v\|^2$$

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.

Lemma. *Each continuously differentiable mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous over convex compact sets.*

8.1.20. 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,

odkud

$$z_x(\pi, 1) = -\frac{F_x(\pi, 1, 0)}{F_z(\pi, 1, 0)} = \frac{1}{\pi + 1},$$

$$z_y(\pi, 1) = -\frac{F_y(\pi, 1, 0)}{F_z(\pi, 1, 0)} = \frac{\pi}{\pi + 1}.$$

$\square$

8.G.4. Having the mapping $F : \mathbb{R}^3 \to \mathbb{R}^2$, $F(x, y, z) = (f(x, y, z), g(x, y, z)) = (e^{x \sin y}, xyz)$, show that the equation $F(x, c_1(x), c_2(x)) = (0, 0)$ defines a curve $c : \mathbb{R} \to \mathbb{R}^2$ 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 8.1.24) 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

$$\begin{pmatrix} (c_1)_x(1) \\ (c_2)_x(1)) \end{pmatrix} = \begin{pmatrix} f_y(1, \pi, 1) & f_z(1, \pi, 1) \\ g_y(1, \pi, 1) & g_z(1, \pi, 1) \end{pmatrix}^{-1} \begin{pmatrix} f_x(1, \pi, 1 \\ g_x(1, \pi, 1) \end{pmatrix} =$$

$$= \begin{pmatrix} -1 & 0 \\ 1 & \pi \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \pi \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ \frac{1}{\pi} & \frac{1}{\pi} \end{pmatrix} \begin{pmatrix} 0 \\ \pi \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

$\square$

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 differential is the one form expressed via the partial derivatives of the outer components, evaluated on the vector of the derivative of the inner component, again given by the product of the one line (the form) and one column (the vector).

Theorem. *Let $F : E_n \to E_m$ and $G : E_m \to E_r$ 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*

$$D^1(G \circ F)(x) = D^1 G(F(x)) \circ D^1 F(x).$$

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.

$$d_v(g \circ F)(x) = \lim_{t \to 0} \frac{1}{t} \big(g(F(x + tv)) - g(F(x)) \big).$$

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 α is a function defined on a neighbourhood of the point $F(x)$ which is continuous and $\lim_{v \to 0} \frac{1}{\|v\|} \alpha(v) = 0$. Substitution into the equality for the directional derivative yields

$$d_v(g \circ F)(x) = \lim_{t \to 0} \frac{1}{t} \Big(dg(F(x))(F(x + tv) - F(x))$$

$$+ \alpha \big(F(x + tv) - F(x) \big) \Big)$$

$$= dg(F(x)) \Big(\lim_{t \to 0} \frac{1}{t} \Big(F(x + tv) - F(x) \Big) \Big)$$

$$+ \lim_{t \to 0} \frac{1}{t} \Big(\alpha \big(F(x + tv) - F(x) \big) \Big)$$

$$= dg(F(x)) \circ D^1 F(x)(v) + 0.$$

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 α.

So the theorem for the particular functions $g_1, \ldots, g_r$ of the mapping G is proved. The theorem in general now follows

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., $ab < a + b$. $\square$

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, $\operatorname{grad} h(x,y,z) = (3x^2, 3y^2, 3z^2)$), we get the system

$$0 = 3x^2 - 2\lambda x,$$
$$0 = 3y^2 - 2\lambda y,$$
$$0 = 3z^2 - 2\lambda z,$$
$$0 = x^2 + y^2 + z^2 - 1$$

consisting of four equations in four variables. Before trying to solve this system, we can estimate how many local constrianed 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

from the definition of matrix multiplication and its links to linear mappings. $\square$

8.1.21. Transformation of coordinates. A mapping $F : E_n \to E_n$ which has an inverse mapping $G : E_n \to E_n$ 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 $x \neq 0$).
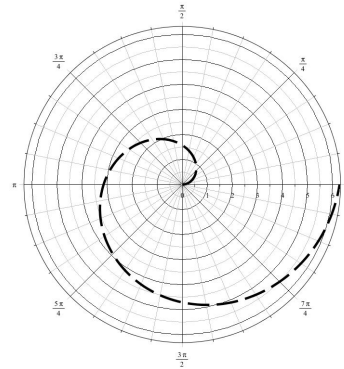
The illustration shows the the "line" $r = \varphi$ drawn in the Cartesian coordinates.

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, φ) in the plane so that the inverse mapping would exist. The Cartesian image of lines in polar coordinates with constant coordinates r or φ is also shown in the illustration above.

We illustrate by an example, usage of the concept of transformation and the theorem about differentiation of composite mappings. The inverse to the above is the transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$ (for instance, on the domain of all points in the first quadrant except for the points having $x = 0$):

$$r = \sqrt{x^2 + y^2}, \; \varphi = \arctan\frac{y}{x}.$$

Consider now the function $g_t : E_2 \to \mathbb{R}$, with free parameter $t \in \mathbb{R}$,

$$g(r, \varphi, t) = \sin(r - t)$$

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 ± 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 η representing the gradient coefficient. This leads to the bigger system

$$0 = 3x^2 - 2\lambda x - \eta,$$
$$0 = 3y^2 - 2\lambda y - \eta,$$
$$0 = 3z^2 - 2\lambda z - \eta,$$
$$0 = x^2 + y^2 + z^2 - 1,$$
$$0 = x + y + z.$$

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. □

8.H.3. Determine whether the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2 y$ 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:

$$2xy = 4kx,$$
$$x^2 = 4ky,$$
$$0 = 2kz.$$

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.
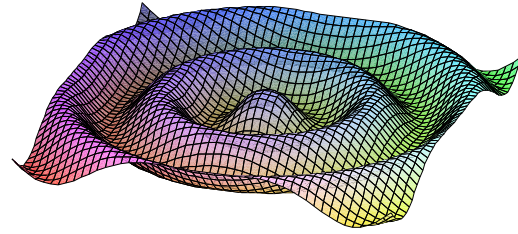
Compute the derivative of this function in Cartesian coordinates. Using the theorem,

$$\frac{\partial g}{\partial x}(x, y, t) = \frac{\partial g}{\partial r}(r, \varphi)\frac{\partial r}{\partial x}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi)\frac{\partial \varphi}{\partial x}(x, y)$$
$$= \cos(\sqrt{x^2 + y^2} - t)\frac{x}{\sqrt{x^2 + y^2}} + 0$$

and, similarly,

$$\frac{\partial g}{\partial y}(x, y, t) = \frac{\partial g}{\partial r}(r, \varphi)\frac{\partial r}{\partial y}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi)\frac{\partial \varphi}{\partial y}(x, y)$$
$$= \cos(\sqrt{x^2 + y^2} - t)\frac{y}{\sqrt{x^2 + y^2}}.$$

8.1.22. 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).

Interpreting this situation for a mapping $E_1 \to E_1$ and linear mappings $\mathbb{R} \to \mathbb{R}$ as their differentials, the nonvanishing is a necessary and sufficient condition for the differential to be invertible as a linear mapping. In this way, a statement is obtained which is valid for all finite-dimensional spaces in general:

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. Ir follows from the above that the first and second stationary points are maxima of the function on the given ellipsoid, while the other two are minima. □

Remark. Note that we have used the variable k instead of λ from the theorem 8.1.28.

8.H.4. Decide whether the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = z - xy^2$ 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

$$kx = -y^2,$$
$$ky = -2xy,$$
$$kz = 1.$$

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. □

8.H.5. Determine whether the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = xyz$, 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:

$$\frac{\partial g}{\partial x} = \lambda\frac{\partial f}{\partial x} : \quad yz = 2\lambda kx,$$
$$\frac{\partial g}{\partial y} = \lambda\frac{\partial f}{\partial y} : \quad xz = 2\lambda ly,$$
$$\frac{\partial g}{\partial z} = \lambda\frac{\partial f}{\partial z} : \quad xy = 2\lambda z.$$

Theorem. *Let $F : E_n \to E_n$ be a differentiable mapping on a neighbourhood of a point $x_0 \in E_n$, and let the Jacobi matrix $D^1F(x_0)$ be invertible.*

Then in some neighbourhood of 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 $D^1F(x_0)$.

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.

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

$$\mathrm{id}_{\mathbb{R}^n} = D^1(F^{-1} \circ F)(x_0) = D^1(F^{-1}) \circ D^1F(x_0),$$

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 $\alpha : \mathbb{R}^n \to 0$ satisfying $\lim_{v\to 0} \frac{1}{\|v\|}\alpha(v) = 0$. 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 $y_0 = F(x_0)$:

$$\lim_{y\to y_0} \frac{1}{\|y - y_0\|}\left(F^{-1}(y) - F^{-1}(y_0) - (D^1F(x_0))^{-1}(y - y_0)\right).$$

Substituting (1) for $y - y_0$ into the latter equality yields

$$\lim_{y\to y_0} \frac{1}{\|y - y_0\|}\bigg(x - x_0$$
$$- (D^1F(x_0))^{-1}(D^1F(x_0)(x - x_0) + \alpha(x - x_0))\bigg)$$
$$= \lim_{y\to y_0} \frac{-1}{\|y - y_0\|}(D^1F(x_0))^{-1}(\alpha(x - x_0))$$
$$= (D^1F(x_0))^{-1}\lim_{y\to y_0} \frac{(-1)}{\|y - y_0\|}(\alpha(x - x_0)),$$

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 α, zero if the values $\|F(x) - F(x_0)\|$ are greater than $C\|x - x_0\|$ for some constant

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 $-\frac{1}{3\sqrt{3kl}}$). $\qquad\square$

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

$$(1) \qquad x \geq 0, \quad y \geq 0, \quad y \leq -x + 3.$$

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

$$f(x,y) = f(x, -x+3) = -5x^2 + 18x - 19, \quad x \in [0,3].$$

$C > 0$. 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.

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 $x_0 = 0 \in R^n$, $y_0 = F(x_0) = 0 \in \mathbb{R}^n$. 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 $G(y) = (D^1 F(0))^{-1}(y)$ gives $D^1(G \circ F)(0) = \mathrm{id}_{\mathbb{R}^n}$ and thus we may assume that

$$D^1 F(0) = \mathrm{id}_{\mathbb{R}^n}.$$

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 δ–neighbourhood U_δ of the origin (in the 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_δ, cf. 8.1.19.

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.

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":

$$[1,1], \quad [0,0], \quad [0,3], \quad [3,0], \quad \left[\tfrac{9}{5}, \tfrac{6}{5}\right]$$

with function values

$$-4, \quad -1, \quad -19, \quad -10, \quad -\tfrac{14}{5},$$

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 $[0,3]$. □

8.H.7. Determine whether the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = y^2 z$ 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 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:

$$0 = 2k + l,$$
$$2yz = k - l,$$
$$y^2 = k + 2l,$$
$$2x + y + z = 1,$$
$$x - y + 2z = 0.$$

The solution of the system is $[x,y,z] = \left[\tfrac{2}{3}, 0, -\tfrac{1}{3}\right]$ and $[x,y,z] = \left[\tfrac{4}{9}, \tfrac{2}{9}, -\tfrac{1}{9}\right]$ (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 $[-1, \tfrac{5}{3}, \tfrac{4}{3}]$ and $[2, -\tfrac{4}{3}, -\tfrac{5}{3}]$. Considering these four points, the function takes on the greatest value at the first marginal point ($f(x,y,z) = \tfrac{100}{27}$), which is its maximum on the given segment, and it takes the least value at the second marginal point ($f(x,y,z) = -\tfrac{80}{27}$), which is thus its minimum there. □

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_δ is not only bijective onto its image, but also that it maps open neighbourhoods of zero onto open neighbourhoods of zero.[2]

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 $a = \tfrac{1}{2} \min_{x \in B} \rho(x)$ 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_δ. 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 $j = 1, \ldots, n$,

$$\frac{\partial h}{\partial x^j}(z) = \sum_{i=1}^{n} 2\big(f_i(z) - y_i\big) \frac{\partial f_i}{\partial x_j}(z) = 0.$$

This is a system of linear equations with variables $\xi_i = f_i(z) - y_i$ and coefficients given by twice the Jacobi matrix $D^1 F(z)$. 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 $i = 1, \ldots, n$, the equality $f_i(z) = y_i$, i.e., $F(z) = y$. □

8.1.23. 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 $F(x,y) = 0$.

---

[2]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.

8.H.8. Find the maximal and minimal values of the polynomial

$$p(x, y) = 4x^3 - 3x - 4y^3 + 9y$$

on the set

$$M = \left\{ [x, y] \in \mathbb{R}^2; \ x^2 + y^2 \leq 1 \right\}.$$

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

$$\left[\tfrac{1}{2}, \tfrac{\sqrt{3}}{2} \right], \quad \left[\tfrac{1}{2}, -\tfrac{\sqrt{3}}{2} \right], \quad \left[-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2} \right], \quad \left[-\tfrac{1}{2}, -\tfrac{\sqrt{3}}{2} \right],$$

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

$$x = \cos t, \quad y = \sin t, \qquad t \in [-\pi, \pi].$$

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

$$\tfrac{1}{\cos^2 t} = 1 + \operatorname{tg}^2 t,$$

which is valid provided both sides are well-defined. We get

$$f'(t) = \cos^3 t \big[-12 \operatorname{tg} t + 3 \left(\operatorname{tg} t + \operatorname{tg}^3 t \right) \\ - 12 \operatorname{tg}^2 t + 9 \left(1 + \operatorname{tg}^2 t \right) \big]$$

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

$$s^3 - s^2 - 3s + 3 =$$
$$0, \quad \text{i. e.} \quad (s - 1) \left(s - \sqrt{3} \right) \left(s + \sqrt{3} \right) = 0.$$

Then, the values

An example of this can be the usual (implicit) definition of straight lines and circles:

$$F(x, y) = ax + by + c = 0, \quad a, b, c \in \mathbb{R}$$
$$F(x, y) = (x - s)^2 + (y - t)^2 - r^2 = 0, \ r > 0.$$

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 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

$$s = 1, \quad s = \sqrt{3}, \quad s = -\sqrt{3}$$

respectively correspond to

$$t \in \{-\tfrac{3}{4}\,\pi, \tfrac{1}{4}\,\pi\}, \quad t \in \{-\tfrac{2}{3}\,\pi, \tfrac{1}{3}\,\pi\}, \quad t \in \{-\tfrac{1}{3}\,\pi, \tfrac{2}{3}\,\pi\}.$$

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

$$f\left(-\tfrac{1}{3}\,\pi\right) = -1 - 3\sqrt{3} < f\left(-\tfrac{3}{4}\,\pi\right) = -3\sqrt{2} <$$
$$f\left(-\tfrac{2}{3}\,\pi\right) = 1 - 3\sqrt{3} < -1,$$
$$f\left(-\pi\right) = f\left(\pi\right) = -1 < 0,$$
$$f\left(\tfrac{2}{3}\,\pi\right) = 1 + 3\sqrt{3} > f\left(\tfrac{1}{4}\,\pi\right) = 3\sqrt{2} > f\left(\tfrac{1}{3}\,\pi\right) =$$
$$-1 + 3\sqrt{3} > 0.$$

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 $\cos\left(-\tfrac{1}{3}\,\pi\right) = \tfrac{1}{2}$, $\sin\left(-\tfrac{1}{3}\,\pi\right) = -\tfrac{\sqrt{3}}{2}$, $\cos\left(\tfrac{2}{3}\,\pi\right) = -\tfrac{1}{2}$, $\sin\left(\tfrac{2}{3}\,\pi\right) = \tfrac{\sqrt{3}}{2}$, 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 $[-1/2, \sqrt{3}/2]$. $\qquad \square$

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 $[-1, 0]$. $\qquad \square$

8.H.10. Compute the local extrema of the function $y = f(x)$ given implicitly by the equation

$$3x^2 + 2xy + x = y^2 + 3y + \tfrac{5}{4}, \quad [x, y] \in$$
$$\mathbb{R}^2 \smallsetminus \left\{ \left[x, x - \tfrac{3}{2}\right]; x \in \mathbb{R} \right\}.$$

$f(x, y) = x^2 + y^2$, 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

$$F(x, c(x)) = F(x, c_1(x), c_2(x)) = 0,$$

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.

8.1.24. The general theorem. Consider a differentiable mapping

$$F = (f_1, \ldots, f_n) : \mathbb{R}^{m+n} \to \mathbb{R}^n.$$

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 $(x_1, \ldots, x_{m+n}) \in \mathbb{R}^{m+n}$ 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.

Solution. In accordance with the theoretical part (see 8.1.24), let us denote

$$F(x, y) = 3x^2 + 2xy + x - y^2 - 3y - \tfrac{5}{4},$$
$$[x, y] \in \mathbb{R}^2 \setminus \left\{ \left[x, x - \tfrac{3}{2}\right] ; x \in \mathbb{R} \right\}$$

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 the 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

$$[x, y] = \left[0, -\tfrac{1}{2}\right], \quad [x, y] = \left[\tfrac{1}{2}, -2\right].$$

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

$$y'' = -\frac{6(-2)-0}{4} > 0 \quad \text{for } [x, y] = \left[0, -\tfrac{1}{2}\right]$$

and

$$y'' = -\frac{6(+2)-0}{4} < 0 \quad \text{for } [x, y] = \left[\tfrac{1}{2}, -2\right].$$

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 $x = 1/2$. $\qquad\square$

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)
$$z_x = f_x(x, y) = \frac{z - 2x - 2}{2z - x - y + 2},$$
$$z_y = f_y(x, y) = \frac{z - 2y - 2}{2z - x - y + 2}.$$

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

$$z_x = 0, \quad \text{i. e.} \quad z - 2x - 2 = 0,$$
$$z_y = 0, \quad \text{i. e.} \quad z - 2y - 2 = 0.$$

Theorem. *Let $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$ 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 differentiable mapping $G : \mathbb{R}^m \to \mathbb{R}^n$ defined on an neighbourhood U of the point $a \in \mathbb{R}^m$ with image $G(U)$ which contains the point 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.

Extend the function F to

$$\tilde{F} : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x, y) \mapsto (x, F(x, y)).$$

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 $\pi : \mathbb{R}^2 \to \mathbb{R}$ 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

$$F(x, f(x)) = F\left(x, \pi(\tilde{F}^{-1}(x, 0))\right)$$

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:

$$F(x, f(x)) = \pi\left(\tilde{F}\left(x, \pi(\tilde{F}^{-1}(x, 0))\right)\right)$$
$$= \pi\left(\tilde{F}(\tilde{F}^{-1}(x, 0))\right) = \pi(x, 0) = 0.$$

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 $(D^1 \tilde{F})^{-1}$.

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

$$[x, y, z] = \left[-3 + \sqrt{6}, -3 + \sqrt{6}, -4 + 2\sqrt{6}\right],$$
$$[x, y, z] = \left[-3 - \sqrt{6}, -3 - \sqrt{6}, -4 - 2\sqrt{6}\right].$$

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 interchangeabel 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:

$$f_{xx}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = f_{yy}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = -\frac{1}{\sqrt{6}},$$

$$f_{xy}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = f_{yx}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = 0,$$

$$f_{xx}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = f_{yy}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = \frac{1}{\sqrt{6}},$$

$$f_{xy}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = f_{yx}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = 0.$$

As for the Hessian, we have

$$Hf\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = \begin{pmatrix} -\frac{1}{\sqrt{6}} & 0 \\ 0 & -\frac{1}{\sqrt{6}} \end{pmatrix},$$

$$Hf\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = \begin{pmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

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 $\left[-3 + \sqrt{6}, -3 + \sqrt{6}\right]$, and there is a strict local minimum at the point $\left[-3 - \sqrt{6}, -3 - \sqrt{6}\right]$. □

8.H.12. Determine the strict local 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. 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 $\mathbb{R}^2 \smallsetminus \{[0,0]\}$, we

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 $\tilde{F} : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ and define $G(x) = \pi \tilde{F}^{-1}(x, 0)$, where $\pi : \mathbb{R}^{m+n} \to \mathbb{R}^n$, $\pi(x, y) = y$. 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 $m + n = 2$ and to divide the matrix

$$(D^1 \tilde{F}^{-1}) = \begin{pmatrix} \mathrm{id}_{\mathbb{R}^m} & 0 \\ D_x^1 F(x,y) & D_y^1 F(x,y) \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

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:

$$\begin{pmatrix} \mathrm{id}_{\mathbb{R}^m} & 0 \\ D_x^1 F(x,y) & D_y^1 F(x,y) \end{pmatrix} \cdot \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} \mathrm{id}_{\mathbb{R}^m} & 0 \\ 0 & \mathrm{id}_{\mathbb{R}^n} \end{pmatrix}.$$

Apparently, it follows that $A = \mathrm{id}_{\mathbb{R}^m}, B = 0, D = (D_y^1 F)^{-1}$, 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. □

8.1.25. The gradient of a function. As seen in the previous paragraph, if F is a continuously differentiable function of n variables, the definition $F(x_1, \ldots, x_n) = b$ 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

$$D^1 F = \left(\frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_n} \right)$$

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.

should look for stationary points, i. e., for the solution of the equations $L_x = 0$, $L_y = 0$ for

$$L(x, y, \lambda) = \tfrac{1}{x} + \tfrac{1}{y} - \lambda \left(\tfrac{1}{x^2} + \tfrac{1}{y^2} - 4 \right), \quad x \neq 0, \ y \neq 0.$$

We thus get the equations

$$-\tfrac{1}{x^2} + \tfrac{2\lambda}{x^3} = 0, \quad -\tfrac{1}{y^2} + \tfrac{2\lambda}{y^3} = 0,$$

which lead to $x = 2\lambda$, $y = 2\lambda$. Considering the set of points in question, the constraint $x = y$ gives the stationary points

$$(1) \qquad \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right], \quad \left[-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right].$$

Now, let us examine the second differential of the function L. We can easily compute that

$$L_{xx} = \tfrac{2}{x^3} - \tfrac{6\lambda}{x^4}, \quad L_{xy} = 0, \quad L_{yy} = \tfrac{2}{y^3} - \tfrac{6\lambda}{y^4}, \quad x \neq 0, \ y \neq 0,$$

whence it follows that

$$d^2 L(x, y) = \left(\tfrac{2}{x^3} - \tfrac{6\lambda}{x^4} \right) dx^2 + \left(\tfrac{2}{y^3} - \tfrac{6\lambda}{y^4} \right) dy^2.$$

Differentiating the constraint $\tfrac{1}{x^2} + \tfrac{1}{y^2} = 4$, we get

$$-\tfrac{2}{x^3} dx - \tfrac{2}{y^3} dy = 0, \quad \text{i. e.} \quad dy^2 = \tfrac{y^6}{x^6} dx^2.$$

Therefore,

$$d^2 L(x, y) = \left[\tfrac{2}{x^3} - \tfrac{6\lambda}{x^4} + \left(\tfrac{2}{y^3} - \tfrac{6\lambda}{y^4} \right) \tfrac{y^6}{x^6} \right] dx^2.$$

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

$$d^2 L \left(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2} \right) = -4\sqrt{2}\, dx^2, \quad d^2 L \left(-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2} \right) = 4\sqrt{2}\, dx^2,$$

which means that there is a strict local maximum of the function f at the point $\left[\sqrt{2}/2, \sqrt{2}/2 \right]$, while at the point $\left[-\sqrt{2}/2, -\sqrt{2}/2 \right]$, there is a strict local minimum. The corresponding values are:

$$(2)$$
$$f \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right) = 2\sqrt{2}, \quad f \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right) = -2\sqrt{2}.$$

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:

$$HL(x, y) = \begin{pmatrix} \tfrac{2}{x^3} - \tfrac{6\lambda}{x^4} & 0 \\ 0 & \tfrac{2}{y^3} - \tfrac{6\lambda}{y^4} \end{pmatrix}.$$

The evaluation

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 $\operatorname{grad} F$.

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 $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and the magnitude of the corresponding directional derivative

$$|d_v F| = \left| \frac{\partial f}{\partial x_1} v_1 + \cdots + \frac{\partial f}{\partial x_n} v_n \right| = \cos \varphi \|D^1 F\| \|v\|,$$

where φ is the angle between the directions of the vector v and the gradient F, see the discussion about angles of vectors and straight lines in the fourth chapter (cf. definition 4.1.18). Thus is observed:

Zde by se hodil obrazek, napr. obr. 2

THE MAXIMAL GROWTH OF A FUNCTION

Proposition. *The gradient $D^1 F = \left(\frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_n} \right)$ 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 $D^1 F$ 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 a non-zero multiple of the gradient, i.e. a multiple of

$$D^1 F = \big(2(x_0 - a), 2(y_0 - b), 2(z_0 - c) \big),$$

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 gradient by 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:

$$HL\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right) = \begin{pmatrix} -2\sqrt{2} & 0 \\ 0 & -2\sqrt{2} \end{pmatrix},$$

$$HL\left(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right) = \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{pmatrix}$$

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. $\square$

8.H.13. Find the global extrema of the function

$$f(x,y) = \tfrac{1}{x} + \tfrac{1}{y}, \quad x \neq 0,\ y \neq 0$$

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

$$-2\sqrt{2} < -\tfrac{1}{10} - 2 \leq f(x,y) \leq \tfrac{1}{10} + 2 < 2\sqrt{2}, \quad \text{if} \quad |x| \geq 10.$$

At the same time, we have (interchanging x and y leads to the same task)

$$-2\sqrt{2} < -\tfrac{1}{10} - 2 \leq f(x,y) \leq \tfrac{1}{10} + 2 < 2\sqrt{2}, \quad \text{if} \quad |y| \geq 10.$$

Hence we can see that the function f must have global extrema on the considered set, and this must happen inside

Theorem. *For a function $F(x_1,\ldots,x_n)$ of n variables and a point $P = (p_1,\ldots,p_n)$ 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*

$$0 = \frac{\partial f}{\partial x_1}(P)(x_1 - p_1) + \cdots + \frac{\partial f}{\partial x_n}(P)(x_n - p_n).$$

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 $\mathbb{R}^n \to \mathbb{R}$ given by multiplying the column of coordinates by the row vector $\mathrm{grad}\,F$). Clearly, the selected point P satisfies the equation. $\square$

8.1.26. 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 φ 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 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

$$I(P) = \frac{\mathrm{grad}\,F \cdot v}{\|\,\mathrm{grad}\,F\|\|v\|}I_0 = \frac{-2x - 2y + 2z}{2\sqrt{3}}I_0$$

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

8.1.27. Tangent and normal spaces. Ideas about tangent and normal lines can be extended to general dimensions. With a mapping $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$, and coordinate functions f_i, one can also consider the n equations for $n + m$ variables

$$f_i(x_1,\ldots,x_{m+n}) = b_i, \quad i = 1,\ldots,n,$$

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 $(x_1,\ldots,x_{m+n}) \in \mathbb{R}^{m+n}$ is (at least locally) the graph of a mapping $G : \mathbb{R}^m \to \mathbb{R}^n$. Technically, it is necessary to have some submatrix in D^1F of the maximal possible rank n.

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

$$f_{\max} = f\left(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2}\right) = 2\sqrt{2}, \quad f_{\min} = f\left(-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2}\right) = -2\sqrt{2}.$$

$\square$

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 $F(x, y, z) = x^2 + y^2 + z^2 - 1$, $G(x, y, z) = x + y + z$ 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

$$L(x, y, z, \lambda_1, \lambda_2) = xyz - \lambda_1\left(x^2 + y^2 + z^2 - 1\right) - \lambda_2\left(x + y + z\right),$$

the equations $L_x = 0$, $L_y = 0$, $L_z = 0$ give

$$yz - 2\lambda_1 x - \lambda_2 = 0,$$

$$xz - 2\lambda_1 y - \lambda_2 = 0,$$

For a fixed choice $b = (b_1, \ldots, b_n)$, 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 $D^1 F$ is the Jacobi matrix of a mapping which implicitly defines a set M and $P = (p_1, \ldots, p_{m+n}) \in M$ is a point such that M is the graph of a mapping in the neighbourhood of the point P,

$$D^1 F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_{m+n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_{m+n}} \end{pmatrix},$$

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:

$$0 = \frac{\partial f_1}{\partial x_1}(P)(x_1 - p_1) + \cdots + \frac{\partial f_1}{\partial x_n}(P)(x_{m+n} - p_{m+n})$$

$$\vdots$$

$$0 = \frac{\partial f_n}{\partial x_1}(P)(x_1 - p_1) + \cdots + \frac{\partial f_n}{\partial x_n}(P)(x_{m+n} - p_{m+n}).$$

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 $f_1, \ldots, f_n$ 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:

$$0 = -\frac{1}{2\sqrt{x^2 + y^2}} 2x \Big|_{x=1, y=0} \cdot (x - 1)$$

$$\quad - \frac{1}{2\sqrt{x^2 + y^2}} 2y \Big|_{x=1, y=0} \cdot y + 1 \cdot (z - 1)$$

$$= -x + z$$

$$0 = -2(x - 1) + y + (z - 1) = -2x + y + z + 1,$$

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 τ and σ.

545

$$xy - 2\lambda_1 z - \lambda_2 = 0,$$

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 we find out that the wanted maximum is

$$f\left(-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right) = f\left(-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right) =$$
$$= f\left(\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right) = \frac{1}{3\sqrt{6}},$$

while the minimum is

$$f\left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right) = f\left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) =$$
$$= f\left(-\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) = -\frac{1}{3\sqrt{6}}.$$

□

8.1.28. 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 $F(x_1, \ldots, x_{m+n}) = 0$. 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 the 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.

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:

<center>LAGRANGE MULTIPLIERS</center>

Theorem. *Let $F = (f_1, \ldots, f_n) : \mathbb{R}^{m+n} \to \mathbb{R}^n$ 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 $D^1 F$ at the point P be n. Then P is a stationary point of a continuously differentiable function $h : \mathbb{R}^{m+n} \to \mathbb{R}$ with respect to the constraints F, if and only if there exist real parameters $\lambda_1, \ldots, \lambda_n$ such that*

$$\text{grad } h = \lambda_1 \text{ grad } f_1 + \cdots + \lambda_n \text{ grad } f_n.$$

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 $x_1, \ldots, x_{m+n}$ of the stationary points P with respect to the constraints, and, on the other hand, the n parameters λ_i in the linear combination. It remains to say that the point P belongs to the implicitly given set M, which represents n

8.H.15. Find the extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2 + y^2 + z^2$, 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:

$$x = k, \ y = k \ z = -k, \quad k \in \mathbb{R}.$$

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 $x + y : \mathbb{R}^3 \to \mathbb{R}$ 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 maximum of the considered function on the given circle at the former point and a minimum at the latter one. □

8.H.17. Find the extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2 + y^2 + z^2$, on the plane $2x + y - z = 1$ and determine their types. ○

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.

8.1.29. Arithmetic mean versus geometric mean. As an example of practical application of the Lagrange multipliers, we prove the inequality

$$\frac{1}{n}(x_1 + \cdots + x_n) \geq \sqrt[n]{x_1 \cdots x_n}$$

for any n positive real numbers $x_1, \ldots, x_n$. Equality occurs if and only if all the x_i's are equal.

Consider the sum $x_1 + \cdots + x_n = c$ as the constraint for a (non-specified) non-negative constant c. We look for the maxima and minima of the function

$$f(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}$$

with respect to the constraint and the assumption $x_1 > 0, \ldots, x_n > 0$.

The normal vector to the hyperplane M defined by the constraint is $(1, \ldots, 1)$. 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:

$$\frac{1}{n} \frac{1}{x_i} \sqrt[n]{x_1 \cdots x_n} = \lambda,$$

for $i = 1, \ldots, n$ and $\lambda \in \mathbb{R}$.

This system has the unique solution $x_1 = \cdots = x_n$ 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 $x_i = \frac{c}{n}$, $i = 1, \ldots, n$, 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.

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

8.H.18. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = xy$ on the circle with radius 1 which is centered at the point $[x_0, y_0] = [0, 1]$. ○

8.H.19. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = xy$ on the circle with radius 1 which is centered at the point $[x_0, y_0] = [2, 0]$. ○

8.H.20. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = xy$ on the circle with radius 1 which is centered at the point $[x_0, y_0] = [2, 0]$. ○

8.H.21. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = xy$ on the ellipse $x^2 + 3y^2 = 1$. ○

8.H.22. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = x^2 y$ on the circle with radius 1 which is centered at the point $[x_0, y_0] = [0, 0]$. ○

8.H.23. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^3 y$ on the circle $x^2 + y^2 = 1$. ○

8.H.24. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = xy$ on th ellipse $2x^2 + 3y^2 = 1$. ○

8.H.25. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = xy$ on the ellipse $x^2 + 2y^2 = 1$. ○

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) $x + y + z \leq 0$.
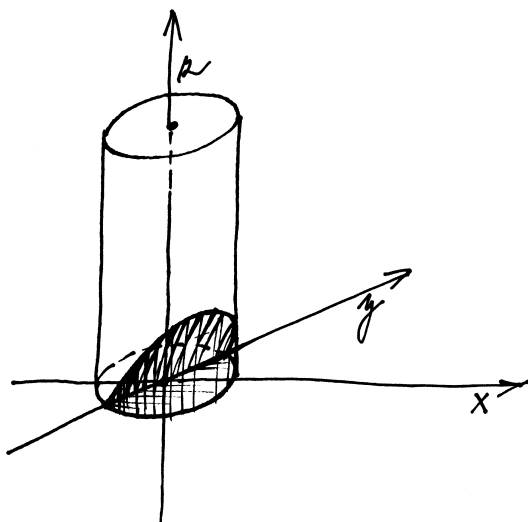
Solution. a) The volume can be calculated with ease using cylindric coordinates. There, the cylinder is determined by

shall see that the approaches coincide for many reasonable functions.

8.2.1. Integrals dependent on parameters. Recall that integrating a function $f(x, y_1, \ldots, y_n)$ of $n + 1$ variables with respect to the single variable x, the result is a function $F(y_1, \ldots, y_n)$ of the remaining variables. Essentially, we proved the following theorem already in 6.3.12 and 6.3.14. 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.

CONTINUITY AND DIFFERENTIATION

Theorem. *Consider a continuous function $f(x, y_1, \ldots, y_n)$ defined for all x from a finite interval $[a, b]$ and for all $(y_1, \ldots, y_n)$ lying in some neighbourhood U of a point $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, and its integral*

$$F(y_1, \ldots, y_n) = \int_a^b f(x, y_1, \ldots, y_n)\, dx.$$

Then function $F(y_1, \ldots, y_n)$ 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

$$\frac{\partial F}{\partial y_j}(c) = \int_a^b \frac{\partial f}{\partial y_j}(x, c_1, \ldots, c_n)\, dx.$$

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 6.3.14, too. □

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 Δx_i which divided the whole interval $[a, b]$. Then, their representatives ξ_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 inequality $r \leq 1$; the half-plane $z \leq x$ by $z \leq r \cos \varphi$, then. Altogether, we get

$$V = \int_0^1 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{r \cos \varphi} r \, \mathrm{d}z \, \mathrm{d}\varphi \, \mathrm{d}r = \frac{2}{3}.$$

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:

$$V = \int_0^1 \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_{-\sqrt{2}r \cos \varphi}^{0} r \, \mathrm{d}z \, \mathrm{d}\varphi \, \mathrm{d}r = \frac{2\sqrt{2}}{3}.$$ 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.
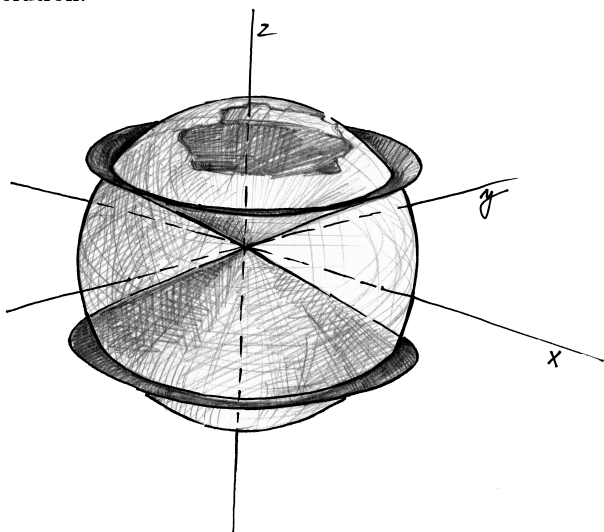
First, we should realize what the examined solid looks like. It is a part of a ball which lies outside a given cone (see the picture).

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

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 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 $\Delta x_{i_1 \ldots i_n}$ 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 $I = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ exists if for every choice of a sequence of divisions Ξ (dividing the multidimensional interval in all variables simultaneously), and the representatives $\xi_{i_1 \ldots i_n}$ of the little mutlidimensional intervals in the partitions, with the norm of the partitions converging to zero, the integral sums

$$S_{\Xi, \xi} = \sum_{i_1 \ldots i_n} f(\xi_{i_1 \ldots i_n}) \Delta x_{i_1 \ldots i_n}$$

always converge to the value

$$S = \int_I f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n,$$

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: control the norms of the divisions used in the definition somewhat roughly.

to halves). We will calculate in spherical coordinates.

$$x = r\cos(\varphi)\sin(\psi),$$
$$y = r\sin(\varphi)\sin(\psi),$$
$$z = r\cos(\psi),$$

$\varphi \in [0, 2\pi)$, $\psi \in [0, \pi)$, $r \in (0, \infty)$.

The Jacobian of this transformation $\mathbb{R}^3 \to \mathbb{R}^3$ 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

$$x^2 + y^2 + z^2 = r^2 \le 1.$$

First, let us find the integration bounds for the variable φ. If we denote by π_φ the projection onto the φ-coordinate in the spherical coordinates ($\pi_\varphi(\varphi, \theta, r) = \varphi$), then the image of the projection π_φ of the solid in question gives the integration bounds for the variable φ. We know that $\pi_\varphi(ball) = [0, 2\pi)$ (the equation $r^2 \le 1$ does not contain the variable φ, 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 φ_0 (fixed from now on), we find the intersection of the solid (ball) and the surface $\varphi = \varphi_0$ and its projection π_ψ on the variable ψ. Similarly like for φ, the variable ψ is not bounded (either by the inequality $r^2 \le 1$ 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 $r^2 \le 1$, $\varphi = \varphi_0$, $\psi = \psi_0$ on the variable r. The only constraint for r is the condition $r^2 \le 1$, 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

$$V_{\text{koule}} = \int_0^1 \int_0^{2\pi} \int_0^{\pi} r^2\sin(\psi)\,\mathrm{d}\psi\,\mathrm{d}\varphi\,\mathrm{d}r = \frac{4}{3}\pi.$$

Now, let us compute the volume of the spherical sector given by $x^2 + y^2 + z^2 \le 1$ and $3x^2 + 3y^2 \ge z^2$. Again, we

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*[3] if and only if its indicator function, defined by

$$\chi_M(x_1, \ldots, x_n) = \begin{cases} 1 & \text{for } (x_1, \ldots, x_n) \in S \\ 0 & \text{for all other points in } \mathbb{R}^n, \end{cases}$$

is Riemann-integrable over $\mathbb{R}^n$.

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

$$\int_M f\,dx_1 \ldots dx_n = \int_{\mathbb{R}^n} \tilde{f}\,dx_1 \ldots dx_n,$$

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:

---

[3] Better to say "measurable via Riemann integration", the measure itself is commonly called the *Peano–Jordan measure* in the literature.

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 φ 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,

$$V = V_{\text{ball}} - V_{\text{sector}} = \frac{2}{3}\pi - \frac{2 - \sqrt{3}}{3}\pi = \frac{\pi}{\sqrt{3}}.$$

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 cylindric 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:

$$V = \frac{2}{3}\pi - \int_0^{2\pi} \int_0^{\frac{1}{2}} \int_0^1 r \, dz \, dr \, d\varphi = \frac{\pi}{\sqrt{3}}.$$

Note that we cannot compute the volume of the solid directly in the cylindric 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.

$$V = V_1 + V_2$$

$$= \int_0^{2\pi} \int_0^{\frac{1}{2}} \int_0^{\sqrt{3}r} r \, dz \, dr \, d\varphi$$

$$+ \int_0^{2\pi} \int_{\frac{1}{2}}^1 \int_0^{\sqrt{1-r^2}} r \, dz \, dr \, d\varphi$$

$$= \frac{\pi}{\sqrt{3}}.$$

$\square$

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

Any linear combination of Riemann-integrable functions $f_i : \mathbb{I} \to \mathbb{R}$, $i = 1, \ldots, k$ (over scalars in $\mathbb{R}$) is again a Riemann-integrable function, and its integral can be computed as follows:

$$\int_I \big(a_1 f_1(x_1, \ldots, x_n) + \ldots + a_k f_k(x_1, \ldots, x_n)\big) dx_1 \ldots dx_n$$

$$= a_1 \int_I f_1(x_1, \ldots, x_n) \, dx_1 \ldots dx_n +$$

$$\cdots + a_k \int_I f_k(x_1, \ldots, x_n) \, dx_1 \ldots dx_n.$$

Let M_1 and M_2 be disjoint Riemann-measurable sets, consider a function $f : M_1 \cup M_2 \to \mathbb{R}$. Then f is Riemann-integrable over both sets M_i if and only if it is integrable over its union, and

$$\int_{M_1 \cup M_2} f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n$$

$$= \int_{M_1} f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n +$$

$$\int_{M_2} f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n.$$

8.2.4. Multiple integrals. Riemann-integrable functions 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 indepdence of the result on our way of describing M. The theorem reveals that the two approaches coincide and there are no unclear points left.

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 $x^2 + y^2 \leq \frac{1}{4}$, $0 \leq z \leq \frac{\sqrt{3}}{2}$.

$$
\begin{aligned}
V &= V_1 + V_2 \\
&= \left(\frac{\pi\sqrt{3}}{8} - \frac{\pi\sqrt{3}}{24} \right) + \left(\pi \int_0^{\frac{\sqrt{3}}{2}} (1 - r^2)\,\mathrm{d}r - \frac{\pi\sqrt{3}}{8} \right) \\
&= \frac{\pi\sqrt{3}}{4} + \frac{\pi}{4\sqrt{3}} = \frac{\pi}{\sqrt{3}}.
\end{aligned}
$$

8.I.3. Calculate the volume of the spherical segment of the ball $x^2 + y^2 + z^2 = 2$ cut by the plane $z = 1$.
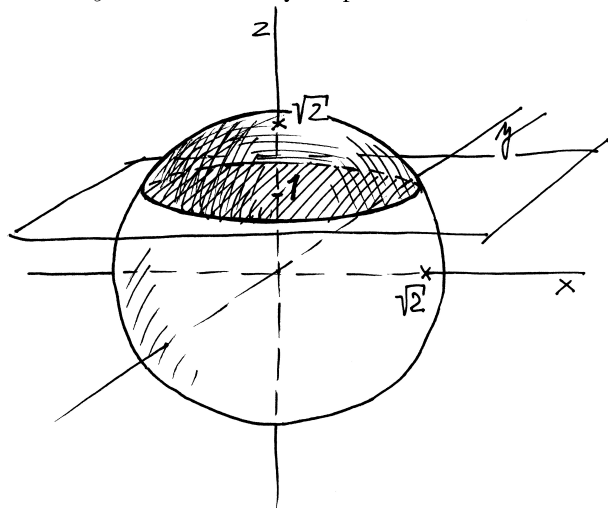
Solution. We iwll compute the integral in spherical coordinates. The segment can be perceived as a spherical sector without the cone (with vertex at the point $[0, 0, 0]$ and the circular base $z = 1$, $x^2 + y^2 = 1$). In these coordinates, the sector is the product of the intervals $[0, \sqrt{2}] \times [0, 2\pi] \times [0, \pi/4]$. We thus integrate in the given bounds, in any order:

$$
\int_0^{2\pi} \int_0^{\sqrt{2}} \int_0^{\frac{\pi}{4}} r^2 \sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{4}{3}(\sqrt{2} - 1)\pi.
$$

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

$$
V_{\text{sector}} - V_{\text{cone}} = \frac{4}{3}(\sqrt{2} - 1) - \frac{1}{3}\pi = \frac{1}{3}\pi(4\sqrt{2} - 5).
$$

Theorem. *Let $M \subset \mathbb{R}^n$ be a bounded set, expressed with the help of continuous functions ψ_i, η_i*

$$
M = \{(x_1, \ldots, x_n);\ x_1 \in [a, b], x_2 \in [\psi_2(x_1), \eta_2(x_1)],
$$
$$
\ldots, x_n \in [\psi_n(x_1, \ldots, x_{n-1}), \eta_n(x_1, \ldots, x_{n-1})]\},
$$

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

$$
\int_M f(x_1, x_2, \ldots, x_n)\,dx_1 \ldots dx_n = \int_a^b \left(\int_{\psi_2(x_1)}^{\eta_2(x_1)} \ldots \right.
$$
$$
\left. \left(\int_{\psi_n(x_1,\ldots,x_{n-1})}^{\eta_n(x_1,\ldots,x_{n-1})} f(x_1, x_2, \ldots, x_n)\,dx_n \right) \ldots dx_2 \right) dx_1
$$

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 Ξ of the interval I with representatives ξ_{ij}.

The corresponding integral sum is

$$
S_{\Xi, \xi} = \sum_{i,j} f(\xi_{ij}) \Delta x_{ij}
$$
$$
= \sum_i \left(\sum_j f(\xi_{ij}) \right) \Delta y_j \, \Delta x_i,
$$

where Δx_{ij} is written for the product of the sizes Δx_i and Δx_j of the intervals which correspond to the choice of the representative ξ_{ij}.

Assume that the work is only with choices of representatives ξ_{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

$$
\sum_i S_i \Delta x_i \to S = \int_a^b \left(\int_{\psi(x)}^{\eta(y)} f(x, y)\,dy \right) dx.
$$

It remains to deal with the case of general choices of representatives of general divisions Ξ. Since f is a continuous function on a compact set, it is uniformly continuous there.

The volume of a general spherical segment with height h in a ball with radius R could be computed similarly:

$$V = V_{\text{sector}} - V_{\text{cone}}$$
$$= \int_0^{2\pi} \int_0^{\arccos\left(\frac{R-h}{R}\right)} \int_0^R r^2 \sin(\theta)\, dr\, d\theta\, d\varphi$$
$$- \frac{1}{3}\pi(2Rh - h^2)(R - h)$$
$$= \frac{1}{3}\pi h^2(3R - h).$$

$\square$

8.I.4. Find the volume of the part of the cylinder $x^2 + z^2 = 16$ which lies inside the cylinder $x^2 + y^2 = 16$.
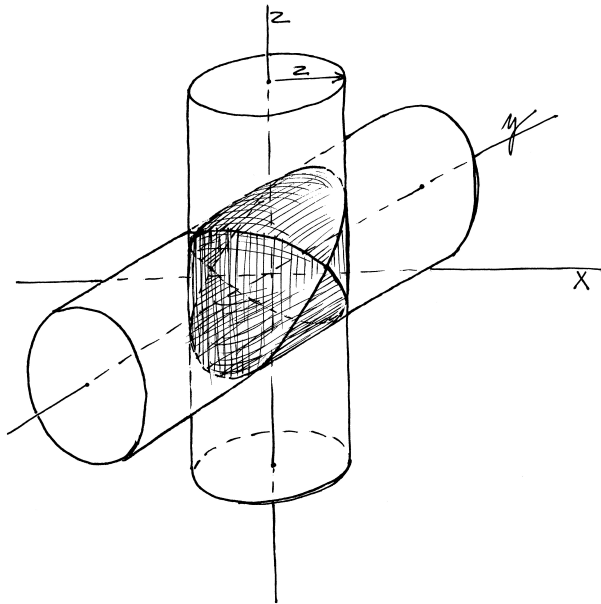
Solution. We will compute the integral in Cartesian coordinates. Since the solid is symmetric, it suffices to integrate over the first octant (interchanging x and $-x$ does not change the equation of the solid; the same holds for y and for z). The part of the solid that lies in the first octant is given 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

$$V = 8\int_0^4 \int_0^{\sqrt{16-x^2}} \frac{4}{\sqrt{16 - x^2}}\, dy\, dx = 128. \qquad \square$$

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.

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 ξ_{ij} differ by no more than ε from the choices used above. 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 $(x_2, \ldots, x_n)$ 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. $\square$

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 any order. We have exploited this behavior already in Chapter 6, cf. 6.3.13. In this way is proved the important corollary:[4]

<div align="center">FUBINI THEOREM</div>

Theorem. *Every continuous function $f(x_1, \ldots, x_n)$ on a multidimensional interval $M = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ is Riemann integrable on M, and its integral*

$$\int_M f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$
$$= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \ldots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

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

---

[4]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.

8.I.5. Find the volume of the part of the cylinder $x^2+y^2 = 4$ bounded by the planes $z = 0$ and $z = x + y + 2$.
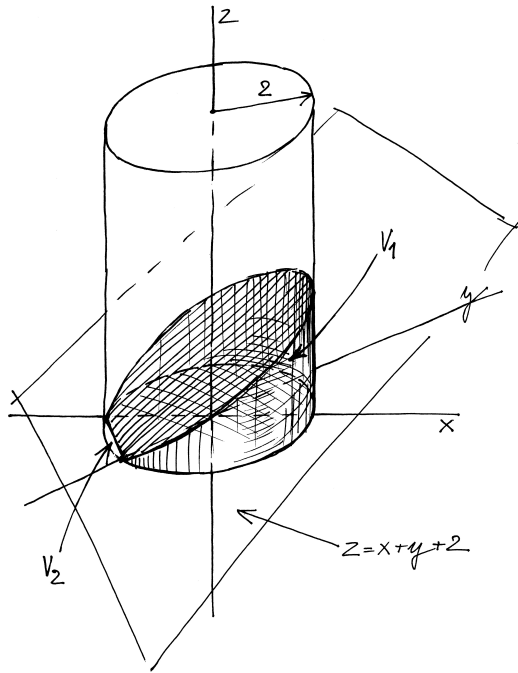
Solution. We will work in cylindric coordinates given by the equations $x = r\cos(\varphi)$, $y = r\sin(\varphi)$, $z = z$. The Jacobian of this transformation is $J = r$. The solid can be divided into two parts: above and below the plane $z = 0$, whose volumes will be denoted by V_1 and V_2, respectively. 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.

$$V_1 - V_{\text{pyramid}} = \int_{-\pi/2}^{\pi}\left(\int_0^2 [r\sin\varphi + r\cos\varphi + 2]r\, dr\right)d\varphi$$
$$= 6\pi + \frac{16}{3},$$
$$V_{\text{pyramid}} = \frac{4}{3}.$$

Further,

$$V_1 - V_2 = \int_{-\pi}^{\pi}\int_0^2 r^2(\sin(\varphi) + \cos(\varphi)) + 2r\, dr\, d\varphi = 8\pi,$$

so $V_1 + V_2 = 4\pi + \frac{40}{3}$. □

Remark. During the calculation, we made use of the fact that integrating a function of two variables over an area in $\mathbb{R}^2$

of an unbounded function is illustrative in this direction:

$$\int_0^1\left(\int_0^1 \frac{x-y}{(x+y)^3}\, dy\right)dx = \frac{1}{2}$$
$$\int_0^1\left(\int_0^1 \frac{x-y}{(x+y)^3}\, dx\right)dy = -\frac{1}{2}.$$

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 Rieman 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

$$\int_{\mathbb{R}^n} f(x)\, dx = \lim_{r\to\infty}\int_{I_r} f(x)\, dx,$$

where $I_r = \{(x_1,\ldots,x_n); |x_j| < r, j = 1,\ldots,n\}$. The resulting limit, if it exists, is bounded by the same constant C. In this case, the Fubini theorem is true in the form

$$\int_{\mathbb{R}^n} f(x)\, dx = \int_{-\infty}^{\infty}\ldots\left(\int_{-\infty}^{\infty} f(x)\, dx_1\right)\ldots dx_n.$$

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 , the Riemann integration of the indicator functions χ_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*[5] 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

_____

[5]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

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 $x^2 + y^2 = 1$.
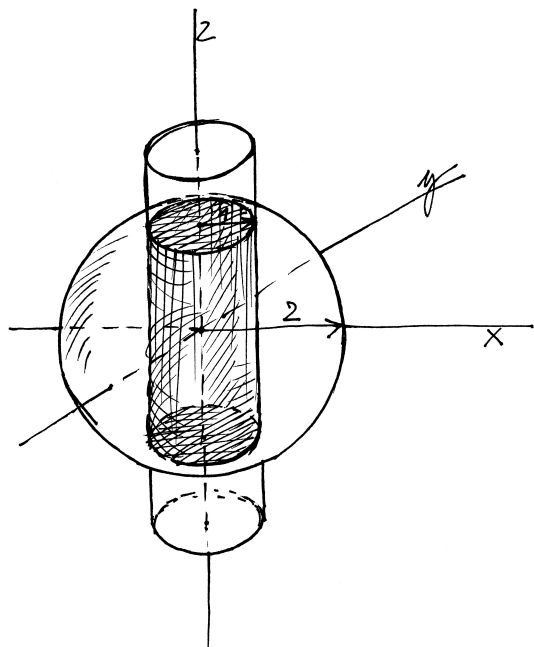
Solution. Thanks to symmetry, it suffices to compute the volume of the part that lies in the first octant. We will integrate in cylindric coordinates given by the equations $x = r\cos(\varphi)$, $y = r\sin(\varphi)$, $z = z$ with Jacobian $J = r$, and it is the space between the plane $z = 0$ and the 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

$$V = 8 \int_0^{\pi/2} \int_0^1 r\sqrt{4 - r^2} \, dr \, d\varphi = \frac{2}{3}(8 - 3\sqrt{3})\pi.$$

$\square$

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 $z = x^2 + y^2$.

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.

$$\|f\|_p = \left(\int_{\mathbb{R}^n} |f(x_1, \ldots, x_n)|^p \, dx_1 \ldots dx_n \right)^{1/p}$$

for any $1 \le p < \infty$. 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 $\| \ \|^p$ 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.

8.2.8. Change of coordinates. When calculating integrals of univariate functions, the "subtitution 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 Δx 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 $\Delta x_{i_1 \ldots i_n}$ in the variables, multiplied by the values of the function at the representative points $\xi_{i_1 \ldots i_n}$. If the coordinates are transformed by means of a mapping $x = G(y)$, not only the function values $f(G(\tilde{\xi}_{i_1 \ldots i_n}))$ are obtained at the
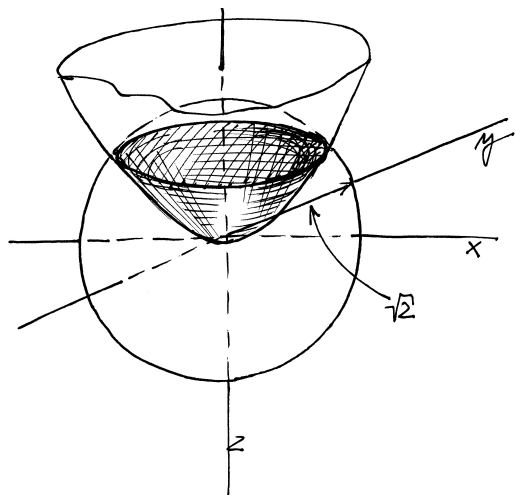
Solution. Once again, we will work in cylindric coordinates:

$$V = \int_0^{2\pi} \int_0^1 \int_{r^2}^{\sqrt{2-r^2}} r \, dz \, dr \, d\varphi = \frac{4\sqrt{2}\pi}{3} - \frac{7\pi}{6}.$$

$\square$

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 $z = 0$.
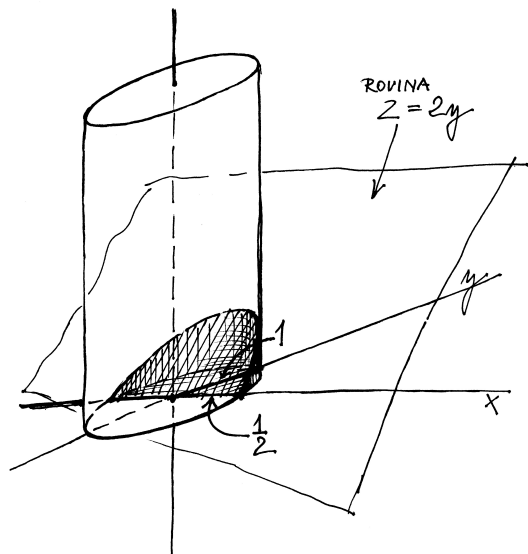
Solution. Thanks to symmetry, it is advantageous to work in the coordinates $x = \frac{1}{2}r\cos(\varphi)$, $y = r\cos(\varphi)$, $z = z$ with Jacobian $J = \frac{1}{2}r$. The equation of the elliptic cylinder in these coordinates is $r^2 = 1$. Thus, the wanted volume is

$$V = \int_0^\pi \int_0^1 r\sin(\varphi)\frac{1}{2}r \, dr \, d\varphi$$

$$= \int_0^\pi \int_0^1 r^2 \sin(\varphi) \, dr \, d\varphi = \int_0^\pi \frac{1}{3}\sin(\varphi) \, d\varphi = \frac{2}{3}.$$

$\square$

representative points $\tilde{\xi}_{i_1\ldots u_n} = G^{-1}(\xi_{i_1\ldots i_n})$ 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 8.1.18. 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 4.1.22).

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.

TRANSFORMATION OF COORDINATES

Theorem. *Let $G(t) : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable and invertible mapping, and write*

$$t = (t_1, \ldots, t_n), \quad x = (x_1, \ldots, x_n) = G(t_1, \ldots, t_n).$$

Further let $M = G(N)$ be a Riemann measurable sets, and $f : M \to \mathbb{R}$ a continuous function. Then, N is also Riemann measurable and

$$\int_M f(x) \, dx_1 \ldots dx_n = \int_N f(G(t)) \big| \det(D^1 G(t)) \big| dt_1 \ldots dt_n.$$

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. 4.1.22 on page 247, 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 $\Psi : \mathbb{R}^n \to \mathbb{R}^n$.

Proposition. *Let $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ 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*

$$\int_M f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n =$$

$$= |\det \Psi| \int_I (f \circ \Psi)(y_1, \ldots, y_n) \, dy_1 \ldots dy_n.$$

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: $\Psi(y_1, \ldots, y_n) = (y_1, \ldots, \alpha y_i, \ldots, y_n)$. In

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 $z = 2$.
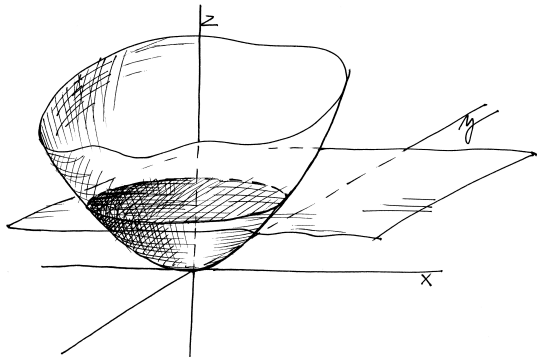
Solution. Similarly to the above problem, we choose "special" coordinates which respect the symmetry of the solid: $x = \frac{1}{\sqrt{2}} r \cos(\varphi)$, $y = r \sin(\varphi)$, $z = z$ with Jacobian $J = \frac{1}{\sqrt{2}} r$. The equation of the paraboloid in these coordinates is $z = r^2$, so the volume of the solid is equal to

$$V = 4 \int_0^{\pi/2} \int_0^{\sqrt{2}} \int_{r^2}^2 \frac{1}{\sqrt{2}} r \, dz \, dr \, d\varphi$$
$$= 2\sqrt{2} \int_0^{\pi/2} \int_0^{\sqrt{2}} 2r - r^3 \, dr \, d\varphi = 2\sqrt{2} \int_0^{\pi/2} d\varphi$$
$$= \sqrt{2}\pi.$$

$\square$

8.I.10. Calculate the volume of the ellipsoid $x^2 + 2y^2 + 3z^2 = 1$.
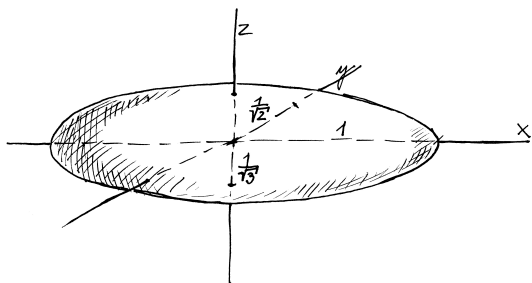
Solution. We will consider the coordinates
$$\begin{aligned} x &= r \cos(\varphi) \sin(\theta), \\ y &= \tfrac{1}{\sqrt{2}} r \sin(\varphi) \sin(\theta), \\ z &= \tfrac{1}{\sqrt{3}} r \cos(\theta). \end{aligned}$$

The corresponding Jacobian is $\frac{1}{\sqrt{6}} r^2 \sin(\theta)$, so the volume is

$$V = \int_0^{2\pi} \int_0^{\pi} \int_0^1 \frac{1}{\sqrt{6}} r^2 \sin(\theta) \, dr \, d\theta \, d\varphi = \frac{4}{3\sqrt{6}}\pi.$$

$\square$

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

this case $|\det \Psi| = |\alpha|$. The second one consists of an exchange of two coordinates, i.e. for given $1 \le i < j \le n$, $\Psi(y_1, \ldots, y_n) = (y_1, \ldots, y_j, \ldots, y_i, \ldots, y_n)$. The determinant of Ψ is -1 in this case. The third type of transformations is of the form $\Psi(y_1, \ldots, y_n) = (y_1, \ldots, y_i + y_j, \ldots, y_j, \ldots, y_n)$, 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 Ψ.

Express the right hand integrals for these three types of Ψ by means of the multiple integrals and Fubini theorem. Write $I = [a_1, b_1] \times \ldots \times [a_n, b_n]$ 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),

$$|\det \Psi| \int_I f(\alpha y_1, y_2, \ldots, y_n) \, dy_1 \ldots dy_n =$$
$$= \alpha \int_{a_n}^{b_n} \ldots \left(\int_{a_1}^{b_1} f(\alpha y_1, y_2, \ldots, y_n) \, dy_1 \right) \ldots dy_n$$
$$= \alpha \, \alpha^{-1} \int_{a_n}^{b_n} \ldots \left(\int_{\alpha a_1}^{\alpha b_1} f(x_1, x_2, \ldots, x_n) \, dx_1 \right) \ldots dx_n$$
$$= \int_{\Psi(I)} f(x_1, x_2, \ldots, x_n) \, dx_1 \ldots dx_n.$$

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:

$$|\det \Psi| \int_I f(y_1 + y_2, y_2, \ldots, y_n) \, dy_1 \ldots dy_n =$$
$$= \int_{a_n}^{b_n} \ldots \left(\int_{a_1}^{b_1} f(y_1 + y_2, y_2, \ldots, y_n) \, dy_1 \right) \ldots dy_n$$
$$= \int_{a_n}^{b_n} \ldots \left(\int_{a_1 + x_2}^{b_1 + x_2} f(x_1, x_2, \ldots, x_n) \, dx_1 \right) \ldots dx_n$$
$$= \int_{\Psi(I)} f(x_1, x_2, \ldots, x_n) \, dx_1 \ldots dx_n.$$

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

557

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

$$
\begin{aligned}
x &= \tfrac{1}{\sqrt{2}} r \cos(\varphi), \\
y &= \tfrac{1}{\sqrt{5}} r \sin(\varphi), \\
z &= z.
\end{aligned}
$$

The determinant of the Jacobian is $\frac{r}{\sqrt{10}}$, so the volume is

$$
V = \int_0^{2\pi} \int_0^1 \int_{r^2}^1 \frac{r}{\sqrt{10}} \, \mathrm{d}z \, \mathrm{d}r \, \mathrm{d}\varphi = \frac{\pi}{2\sqrt{10}}.
$$

$\square$

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 cylindric coordinates,

$$
V = \int_0^{\pi/2} \int_0^3 \int_0^{\frac{r^2}{3}\cos^2(\varphi)} r \, \mathrm{d}x \, \mathrm{d}r \, \mathrm{d}\varphi = \frac{27}{16}\pi.
$$

$\square$

8.I.14. Find the volume of the solid in $\mathbb{R}^3$ which is bounded by the cone part $2x^2 + y^2 = (z-2)^2$, $z \ge 2$ and the paraboloid $2x^2 + y^2 = 8 - z$.
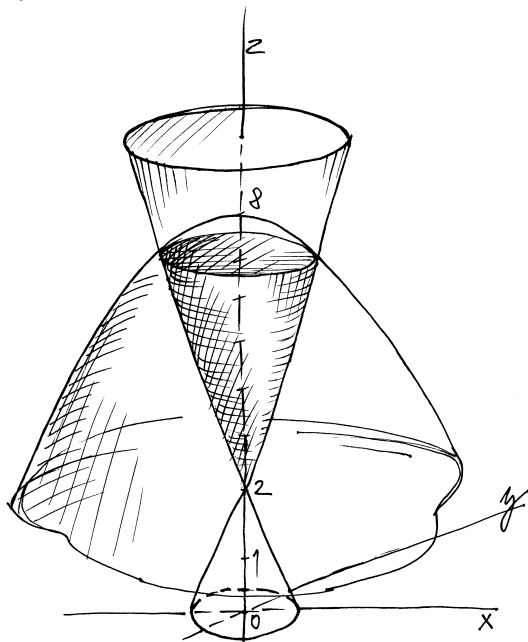
well-defined values of the upper and lower integrals found in this way. If this is done for the indicator function χ_M of a fixed set M, the inner and outer Riemann measure of the set M is obtained. Evidently, the inner measure is the supremum of the areas given by the the (finite) sums of the volumes 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 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 ∂M, i.e. the set of all boundary points of M. For any partition Ξ of I from the definition of the Riemann integral of χ_M, each of the intervals $I_{i_1 \dots i_n}$ with non-trivial intersection with ∂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 $G : M \subset \mathbb{R}^n \to R^n$ 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 ∂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 ∂M.

Next, consider a partition Ξ of an interval I containing ∂M and a fixed tiny interval J in a partition including a point $t \in \partial M$. Write $R = G(t) + D^1 G(t)(J - t)$. 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 $D^1 G$, for each $\varepsilon > 0$ there is a bound δ for the norm of a partition for which

$$
G(J) \subset G(t) + (1+\varepsilon) D^1 G(t)(J - t)
$$

Solution. First of all, we find the intersection of the given surfaces:

$$(z-2)^2 = -z + 8, \; z \geq 2;$$

therefore, $z = 4$, and the equation of the intersection is $2x^2 + y = 4$. 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 φ 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 φ_0) the interval $[0, 2]$. Having fixed both r_0 and φ_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

$$V = \int_0^{2\pi} \int_0^2 \int_{r+2}^{8-r^2} \frac{r}{\sqrt{2}} \, dz \, dr \, d\varphi = \frac{16\sqrt{2}}{3} \pi.$$

$\square$

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 cylindric coordinates,

$$V = \int_0^{2\pi} \int_0^2 \int_0^{8 - \frac{r^2}{2}} r \, dx \, dr \, d\varphi = 28\pi.$$

$\square$

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:

$$x_t = \iiint_T x \, dx \, dy \, dz,$$

$$y_t = \iiint_T y \, dx \, dy \, dz,$$

$$z_t = \iiint_T z \, dx \, dy \, dz.$$

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 $\mathbb{R}^2$.

can be guaranteed. The entire image of J lies inside a slightly enlarged linear image of J by the derivative. Now, the outer measure α of the image $G(J)$ satisfies:

$$\alpha \leq (1 + \varepsilon)^n \operatorname{vol}_n R = (1 + \varepsilon)^n |\det G(t)| \operatorname{vol}_n J.$$

If μ is the upper Riemann sum for the measure of ∂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, for the same ε, the norm of the partition is bounded, 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. $\square$

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

$$g(t_1, \ldots, t_n) = f(G(t_1, \ldots, t_n)),$$

and choose a sufficiently large n-dimensional interval I containing N and its partition Ξ. 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 $G(I_{i_1 \ldots i_n})$ of the intervals from the partition are again Riemann measurable sets. For each small part $I_{i_1 \ldots i_n}$ of the partition Ξ, the integral of f over $J_{i_1 \ldots i_n} = G(I_{i_1 \ldots i_n})$ certainly exists, too.

Further, if the center $t_{i_1 \ldots i_n}$ of the interval $I_{i_1 \ldots i_n}$ is fixed, then the linear image of this interval

$$R_{i_1 \ldots i_n} = G(t_{i_1 \ldots i_n}) + D^1 G(t_{i_1 \ldots i_n})(I_{i_1 \ldots i_n} - t_{i_1 \ldots i_n}),$$

is obtained. This is an n-dimensional parallelepiped (note that the interval is shifted to the origin with 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 $J_{i_1, \ldots i_n}$. 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

$$G(t_{i_1 \ldots i_n}) + (1 + \varepsilon) D^1 G(t_1, \ldots t_n)(I_{i_1 \ldots i_n}) \supset J_{i_1 \ldots i_k}.$$

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

$$S = \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} dy\, dx = \frac{1}{\sqrt{6}} \int_0^1 \int_0^{\sqrt{1-x^2}} dy'\, dx'$$
$$= \frac{\pi}{4\sqrt{6}}.$$

The other integrals we need can be computed directly in Cartesian coordinates x and y:

$$T_x = \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} x\, dy\, dx = \int_0^{\frac{1}{\sqrt{3}}} x\sqrt{\frac{1-3x^2}{2}}\, dx =$$
$$= \frac{1}{2} \int_0^{\frac{1}{3}} \sqrt{\frac{1-3t}{2}}\, dt = \frac{\sqrt{2}}{18},$$

$$T_y = \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} y\, dy\, dx = \frac{1}{2} \int_0^{\frac{1}{\sqrt{3}}} \frac{1-3x^2}{2}\, dx =$$
$$= \frac{1}{4} \int_0^{\frac{1}{\sqrt{3}}} (1-3x^2)\, dx = \frac{\sqrt{3}}{18}.$$

Therefore, the coordinates of the centroid are $\left[\frac{4\sqrt{3}}{9\pi}, \frac{2\sqrt{2}}{\pi}\right]$. □

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 cylindric coordinates that

$$V = 4 \int_0^{\pi/2} \int_0^r \int_{\frac{h}{r}\rho}^h \rho\, dz\, d\rho\, d\varphi = \frac{1}{3}\pi h r^2.$$

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. □

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

However, then the n-dimensional volumes also satisfy

$$\text{vol}_n(J_{i_1\dots i_n}) \le (1+\varepsilon)^n \text{vol}_n(R_{i_1\dots i_n})$$
$$= (1+\varepsilon)^n \big|\det G(t_{i_1\dots i_k})\big| \text{vol}_n(I_{i_1\dots i_n}).$$

Now, it is possible to estimate the entire integral:

$$\int_M f(x_1, \dots, x_n)\, dx_1 \dots dx_n =$$
$$= \sum_{i_1\dots i_n} \int_{J_{i_1\dots in}} f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$
$$\le \sum_{i_1\dots i_n} \big(\sup_{t\in I_{i_1\dots in}} g(t)\big) \text{vol}_n(J_{i_1\dots i_n})$$
$$\le (1+\varepsilon)^n \sum_{i_1\dots i_n} \big(\sup_{t\in I_{i_1\dots in}} g(t)\big)\big|\det G(t_{i_1\dots i_k})\big| \text{vol}_n(I_{i_1\dots i_n}).$$

If ε 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 of $g(t)|\det G(t)|$ is obtained. Instead of the desired equality, the inequality:

$$\int_M f(x)\, dx_1 \dots dx_n \le \int_N f(G(t))\big|\det(D^1 G(t))\big|\, dt_1 \dots dt_n$$

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:

$$\int_N g(t)\big|\det(D^1 G(t))\big|\, dt_1 \dots dt_n$$
$$\le \int_M f(x)\big|\det(D^1 G(G^{-1}(x)))\big|$$
$$\big|\det(D^1 G^{-1}(x))\big|\, dx_1 \dots dx_n$$
$$= \int_M f(x)\, dx_1 \dots dx_n.$$

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).

determine the bounds of the integral that corresponds to the volume of the examined solid:

$$V = \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2\cos\varphi} \int_0^{2r^2} r \, \mathrm{d}z \, \mathrm{d}r \, \mathrm{d}\varphi$$

$$= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2\cos\varphi} 2r^3 \, \mathrm{d}r \, \mathrm{d}\varphi$$

$$= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} 8\cos^4 \pi\varphi = 3\pi,$$

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:

$$x_T = \frac{1}{V} \int \int \int_B x \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$$

$$= \frac{1}{V} \int_0^{2r^2} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2\cos\varphi} r^2 \cos\varphi \, \mathrm{d}z \, \mathrm{d}r \, \mathrm{d}\varphi$$

$$= \frac{1}{V} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2\cos\varphi} 2r^4 \cos\varphi \, \mathrm{d}r \, \mathrm{d}ph$$

$$= \frac{1}{V} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} -\frac{64}{5}\cos^6 \varphi \, \mathrm{d}\varphi = -\frac{4}{3},$$

where the last integral was computed by 6.2.6 again.

Analogously for the z-coordinate of the centroid:

$$z_T = \frac{1}{V} \int_0^{2r^2} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2\cos\varphi} zr \cos\varphi \, \mathrm{d}z \, \mathrm{d}r \, \mathrm{d}\varphi = \frac{20}{9}.$$

The coordinates of the centroid are thus $[-\frac{4}{3}, 0, \frac{20}{9}]$. □

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 cylindric 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
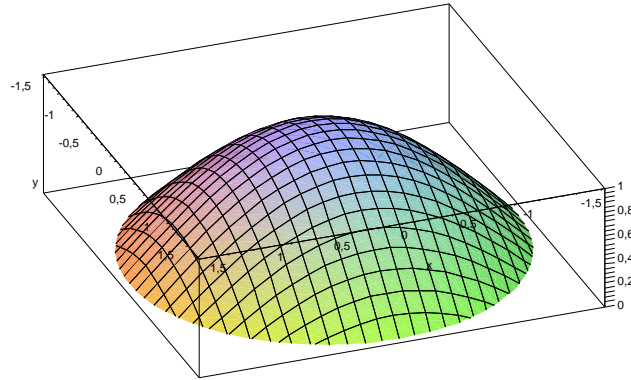
First, determine the Jacobi matrix of the transformation $x = r\cos\theta$, $y = r\sin\theta$

$$D^1 G = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

Hence, the determinant of this matrix is equal to

$$\det D^1 G(r, \theta) = r(\sin^2\theta + \cos^2\theta) = r.$$

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:

$$\int_S dxdy = \int_0^{2\pi} \int_0^R r \, drd\theta = \int_0^R 2\pi r \, dr = \pi R^2.$$

The integration of the function f is very similar, using multiple integration and integration by parts:

$$\int_S f(x, y) \, dxdy = \int_0^{2\pi} \int_0^{\pi/2} r\cos r \, drd\theta = \pi^2 - 2\pi.$$

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.

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.

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 $[0, 0, \frac{3}{2}]$. $\qquad\square$

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 cylindric coordinates, which evaluates as follows:

$$V = \int_0^{2\pi} \int_0^1 \int_{r+2}^{4-r^2} r \, dz \, dr \, d\varphi = \frac{5}{6}\pi.$$

$\qquad\square$

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.

$$V = \int_0^{2\pi} \int_0^1 \int_{r^2}^{2-r} r \, dz \, dr \, d\varphi = \frac{5}{6}\pi.$$

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. $\qquad\square$

8.I.23. Find the centroid of the surface bounded by the parabola $y = 4 - x^2$ and the line $y = 0$. $\qquad\bigcirc$

8.I.24. Find the centroid of the circular sector corresponding to the angle of $60°$ that was cut out of a disc with radius 1. $\bigcirc$

8.I.25. Find the centroid of the semidisc $x^2 + y^2 = 1$, $y \geq 0$. $\qquad\bigcirc$

8.I.26. Find the centroid of the circular sector corresponding to the angle of $120°$ that was cut out of a disc with radius 1. $\bigcirc$

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 $(x - 1)^2 + y^2 \leq 1$. $\bigcirc$

8.I.28. Find the volume of the solid in $\mathbb{R}^3$ which is given by the inequalities $z \geq 0$, $z - y \leq 0$. $\qquad\bigcirc$

8.I.29. Find the volume of the solid bounded by the surface

$$3x^2 + 2y^2 + 3z^2 + 2xy - 2yz - 4xz = 1.$$

$\bigcirc$

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 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) $$y'(t) = r \cdot y(t) + s.$$

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) $$p'(t) = p(t) \left(-\frac{r}{K} p(t) + r \right).$$

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

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 $x \geq 1$. ◯

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

$$P = \int_S \sqrt{1 + f_x^2 + f_y^2} \, \mathrm{d}x \, \mathrm{d}y.$$

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

$$f_x = \frac{x}{x^2 + y^2}, \ f_y = \frac{y}{x^2 + y^2},$$

so the area is expressed by the integral

$$\iint_K \sqrt{1 + f_x^2 + f_y^2} \, \mathrm{d}x \, \mathrm{d}y = \iint_K \sqrt{2} \, \mathrm{d}x \, \mathrm{d}y =$$
$$= \sqrt{2} \int_0^\pi \int_0^{\sin \pi} r \, \mathrm{d}r \, \mathrm{d}\varphi = \frac{\sqrt{2}}{2} \int_0^\pi \sin^2 \varphi$$
$$= \frac{\sqrt{2}\pi}{4}.$$

□

8.I.32. Find the area of the parabola $z = x^2 + y^2$ over the disc $x^2 + y^2 \leq 4$. ◯

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 $y \geq x$. ◯

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

$$I = c\varepsilon_0 \frac{1}{\tau} \int_0^\tau E_y^2 \, \mathrm{d}t,$$

where c is the speed of light and ε_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

is easily verified that the following function is a solution for every constant C:

$$p(t) = \frac{K}{1 + CK \, \mathrm{e}^{-rt}}.$$

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.
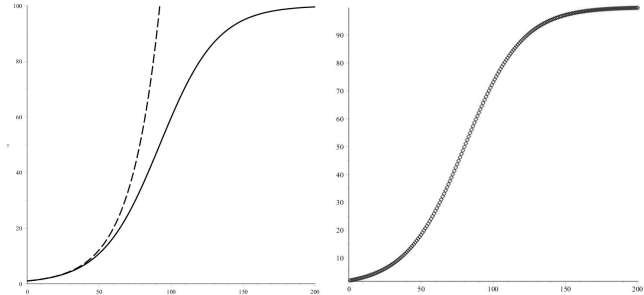
In particular, both versions of this logistic model yield quite similar results. For example, the left hand illustration also contains the dashed line of the graph of the solution of the equation (1) with the same constant r and initial condition (i.e. the Mathusian model of growth).

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 $F : \mathbb{R}^3 \to \mathbb{R}$ 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 $f : \mathbb{R}^2 \to \mathbb{R}$, 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).

maximal amplitude of the wave, ω is the angular frequency, and for any fixed t, the so-called wave length λ is the prime period. The number k then represents the speed $k = \frac{2\pi}{\lambda}$ at which the wave propagates. We have

$$
\begin{aligned}
I &= c\varepsilon_0 \frac{1}{\tau} \int_0^\tau E_y^2 \, \mathrm{d}t = c\varepsilon_0 \frac{1}{\tau} \int_0^\tau A^2 \cos^2(\omega t - k\,x) \, \mathrm{d}t \\
&= c\varepsilon_0 A^2 \frac{1}{\tau} \int_0^\tau \frac{1 + \cos(2(\omega t - k\,x))}{2} \, \mathrm{d}t \\
&= \frac{1}{2} c\varepsilon_0 A^2 \frac{1}{\tau} \big[t + \frac{\sin(2(\omega t - k\,x))}{2\omega} \big]_0^\tau \\
&= \frac{1}{2} c\varepsilon_0 A^2 \frac{1}{\tau} \big(\tau + \frac{\sin(2(\omega\tau - k\,x)) - \sin(2(-k\,x))}{2\omega} \big) \\
&= \frac{1}{2} c\varepsilon_0 A^2 \big(1 + \frac{\sin(2(\omega\tau - k\,x)) - \sin(2(-k\,x))}{2\omega\tau} \big) \\
&\doteq \frac{1}{2} c\varepsilon_0 A^2
\end{aligned}
$$

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 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 (ξ, η) 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 (ξ, η, z):
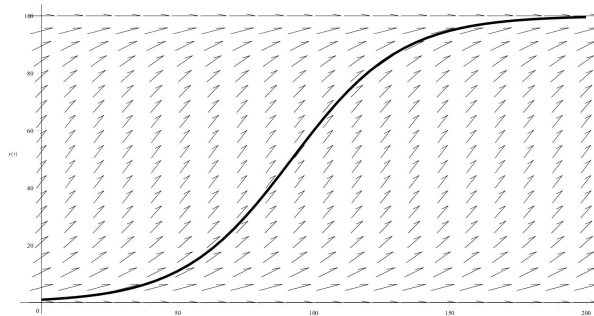
Such illustrations should invoke the idea that differential equations define a "flow" in the plane, and each choice of the initial value $(t_0, y(t_0))$ should correspond to a unique flow-line expressing the movement of the initial point in the time t. It can be anticipated intuitively that for reasonably behaved functions $f(t, y)$ in the equations $y' = f(t, y)$, there is a unique solution for all initial conditions.

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 , is to transforms the problem to ordinary integration, which usually leads to an implicit description of the solution.

<div align="center">EQUATIONS WITH SEPARATED VARIABLES</div>

Consider a differential equation in the form

$$(1) \qquad y' = f(t) \cdot g(y)$$

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

$$(2) \qquad F(t) + C = G(y)$$

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 $y(t_0) = 0$.

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

$$(1) \qquad y' = a(t)y + b(t)$$

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

$$\psi(\xi, \eta) = A \iint_{\mathbb{R}^2} s(x, y) e^{-ik(\xi x + \eta y)}\,dx\,dy$$

$$\psi(\xi, \eta) = A \int_{-p/2}^{p/2} \int_{-q/2}^{q/2} e^{-ik(\xi x + \eta y)}\,dy\,dx$$

$$\psi(\xi, \eta) = A \int_{-p/2}^{p/2} e^{-ik\xi x}\,dx \int_{-q/2}^{q/2} e^{-ik\eta y}\,dy$$

$$= A \left[\frac{e^{-ik\xi x}}{-ik\xi}\right]_{-p/2}^{p/2} \left[\frac{e^{-ik\eta y}}{-ik\eta}\right]_{-q/2}^{q/2}$$

$$= A \frac{2\sin(k\,\xi p/2)}{k\xi} \frac{2\sin(k\,\eta q/2)}{k\eta}$$

$$= Apq \frac{\sin(k\,\xi p/2)}{k\xi p/2} \frac{\sin(k\,\eta q/2)}{k\eta q/2}$$

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:

And the diffraction we are describing:
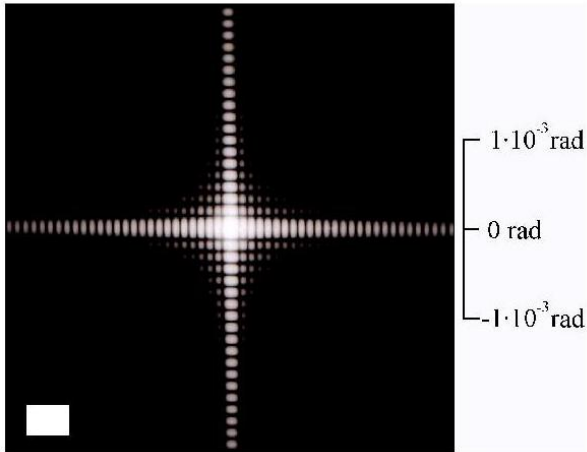
Since $\lim_{x\to 0} \frac{\sin x}{x} = 1$, the intensity at the middle of the image is directly proportional to $I_0 = A^2 p^2 q^2$. The Fourier transform can be easily scrutinized if we aim a laser pointer through a subtle opening between the thumb and the index finger; it will be the image of the 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

$$y' = \frac{\sqrt{1-y^2}}{\cos^2 x}\left(1 + \cos^2 x\right).$$

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) $\qquad \arcsin y = \operatorname{tg} x + x + C, \quad C \in \mathbb{R}.$

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,
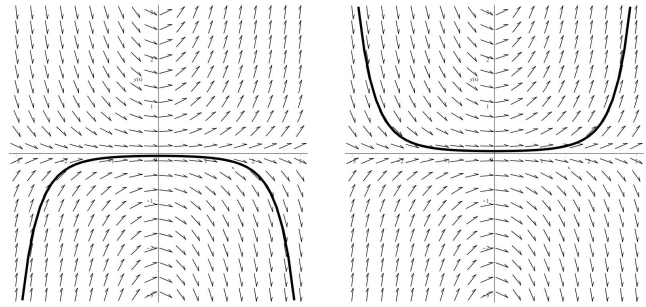
both in the upper bound and a free parameter in the integrand, cf. 6.3.14).

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 $y(t) = c\,F(t, t_0)$ 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 **??** on the page **??**, which could serve as inspiration, too.

As an example, the equation

$$y' = 1 - x\,y,$$

can be solved directly, this time encountering stable behaviour, visible in the following illustration.
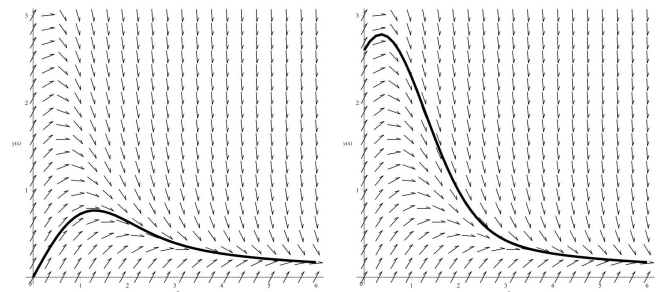
8.3.5. Transformation of coordinates. The illustrations suggest that differential equations can be perceived as geometric objects (the "directional field of the arrows"), so the solution can be found by conveniently chosen coordinates. We return to this point of view later. Here are three simple examples 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 $z = \frac{y}{t}$ and assuming that $t \neq 0$, then by the chain rule,

$$z'(t) = \frac{1}{t^2}\left(t\,y'(t) - y(t)\right) = \frac{1}{t}(f(z) - z),$$

which is an equation with separated variables.

Other examples are the *Bernoulli differential equations*, which are of the form

$$y'(t) = f(t)y(t) + g(t)y(t)^n,$$

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

$$\frac{dz}{dy} \cdot \frac{dy}{dx} = \frac{dz}{dx}, \qquad \frac{1}{\frac{dy}{dx}} = \frac{dx}{dy}.$$

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

$$(2) \qquad y = \sin\left(\operatorname{tg} x + x + C\right), \quad C \in \mathbb{R}.$$

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. □

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

$$\frac{dy}{dx} = (2 - y)\operatorname{tg}\, x,$$

$$-\frac{dy}{y - 2} = \frac{\sin x}{\cos x}\, dx,$$

$$-\ln|y - 2| = -\ln|\cos x| - \ln|C|, \quad C \neq 0.$$

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

where $n \neq 0, 1$. The choice of the transformation $z = y^{1-n}$ leads to the equation

$$z'(t) = (1 - n)y(t)^{-n}(f(t)y(t) + g(t)y^n)$$
$$= (1 - n)f(t)z(t) + (1 - n)g(t),$$

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

$$y'(t) = f(t)y(t) + g(t)y(t)^2 + h(t).$$

This equation can also be transformed to a linear equation provided that a particular solution $x(t)$ can be guessed. Then, use the transformation

$$z(t) = \frac{1}{y(t) - x(t)}.$$

Verify by yourselves that this transformation leads to the equation

$$z'(t) = -\big(f(t) + 2x(t)g(t)\big)z(t) - g(t).$$

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.[6] 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 $f(t, y) : \mathbb{R}^2 \to \mathbb{R}$ 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 $y(t) : I \to \mathbb{R}$ which is the solution of the equation $y' = f(t, y)$ on the interval I.*

---

[6]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.

$$\ln|y-2| = \ln|C\cos x|, \quad C \neq 0,$$
$$|y-2| = |C\cos x|, \quad C \neq 0,$$
$$y-2 = C\cos x, \quad C \neq 0,$$

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

$$y = 2 + C\cos x, \quad C \in \mathbb{R}. \qquad \square$$

8.J.3. Find the solution of the differential equation

$$(1 + \mathrm{e}^x)\, yy' = \mathrm{e}^x$$

which satisfies the initial condition $y(0) = 1$.

Solution. If the functions $f : (a,b) \to \mathbb{R}$ and $g : (c,d) \to \mathbb{R}$ 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

$$\int\limits_{y_0}^{y(x)} \frac{dt}{g(t)} = \int\limits_{x_0}^{x} f(t)\, dt.$$

In practical problems, we first find all solutions of the equation and then select the one which satisfies the initial condition.

Let us compute:

$$(1 + \mathrm{e}^x)\, y\, dy/dx = \mathrm{e}^x,$$
$$y\, dy = \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}\, dx,$$
$$\frac{y^2}{2} = \ln(1 + \mathrm{e}^x) + \ln|C|, \quad C \neq 0,$$
$$\frac{y^2}{2} = \ln(C\,[1 + \mathrm{e}^x]), \quad C > 0.$$

The substitution $y = 1$, $x = 0$ then gives

$$\tfrac{1}{2} = \ln(C \cdot 2), \quad \text{i. e.} \quad C = \tfrac{\sqrt{\mathrm{e}}}{2}.$$

We have thus found the solution

$$\frac{y^2}{2} = \ln\left(\frac{\sqrt{\mathrm{e}}}{2}\,[1 + \mathrm{e}^x]\right),$$

i. e.,

$$y = \sqrt{2\ln\left(\frac{\sqrt{\mathrm{e}}}{2}\,[1 + \mathrm{e}^x]\right)}$$

PROOF. If a differentiable function $y(t)$ is a solution of an equation satisfying the initial condition $y(t_0) = t_0$, 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.

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)

$$\left|\big(L(y) - L(z)\big)(t)\right| = \left|\int_{t_0}^{t} f(s, y(s)) - f(s, z(s))\, ds\right|$$
$$\leq \int_{t_0}^{t} |f(s, y(s)) - f(s, z(s))|\, ds$$
$$\leq C \int_{t_0}^{t} |y(s) - z(s)|\, ds$$
$$\leq C\big(\max_{t_0 \leq s \leq t} |y(s) - z(s)|\big)|t - t_0|$$
$$= D\big(\max_{t_0 \leq s \leq t} |y(s) - z(s)|\big),$$

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 492. At the same time, the operator

on a neighborhood of the point $[0, 1]$ where $y > 0$. $\qquad\square$

8.J.4. Find the solution of the differential equation
$$y' = \tfrac{y^2+1}{x+1}$$
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
$$\arctan 1 = \ln|1| + C, \quad \text{i. e.,} \quad C = \tfrac{\pi}{4}.$$

Therefore, the solution of the given initial problem is the function
$$y(x) = \operatorname{tg}\left(\ln|x + 1| + \tfrac{\pi}{4}\right)$$
on a neighborhood of the point $[0, 1]$. $\qquad\square$

8.J.5. Solve

(1)
$$y' = \frac{x + y + 1}{2x + 2y - 1}.$$

Solution. Let a function $f : (a, b) \times (c, d) \to \mathbb{R}$ 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
$$z' - 1 = \frac{z + 1}{2z - 1},$$
$$\frac{dz}{dx} = \frac{z + 1}{2z - 1} + 1,$$
$$\frac{dz}{dx} = \frac{3z}{2z - 1},$$
$$\left(\frac{2}{3} - \frac{1}{3z}\right) dz = 1\, dx,$$
$$\frac{2}{3} z - \frac{1}{3} \ln|z| = x + C, \quad C \in \mathbb{R},$$
or
$$\tfrac{2}{3} z - \tfrac{1}{3} \ln|Cz| = x, \quad C \neq 0.$$

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 ε and further shrinking δ, implies
$$\max_{t \in J} |L(y)(t) - y_0| < \varepsilon.$$

Finally, the above estimate for $\|L(y) - L(z)\|$ shows that if δ 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 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 499, 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 $y_1 = \lim_{t \to t_{1-}} y(t)$ 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. $\qquad\square$

Now, we must get back to the original variable y in one of these forms. The general solution can be written as

$$\tfrac{2}{3}\,x + \tfrac{2}{3}\,y - \tfrac{1}{3}\ln|x+y| = x + C, \quad C \in \mathbb{R},$$

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)

$$u' = \tfrac{1}{x}\left(f(u) - u\right), \quad \text{i. e.} \quad u'x + u = f(u).$$

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

$$P(x, y)\,dx + Q(x, y)\,dy = 0$$

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

$$y' = \tfrac{y}{x}\ln\tfrac{y}{x}.$$

The substitution $u = y/x$ then leads to

$$u'x + u = u\,\ln u,$$
$$\frac{du}{dx}\,x = u\,(\ln u - 1),$$
$$\frac{du}{u\,(\ln u - 1)} = \frac{dx}{x},$$

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).

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):
$$y_0(t) = y_0, \quad y_{n+1}(t) = L(y_n), \ n = 1, \ldots.$$
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.)

where $u(\ln u - 1) \neq 0$. Using another substitution, namely $t = \ln u - 1$, we can integrate

$$\int \frac{du}{u(\ln u - 1)} = \int \frac{dx}{x},$$

$$\int \frac{dt}{t} = \int \frac{dx}{x},$$

$$\ln|t| = \ln|x| + \ln|C|, \quad C \neq 0,$$

$$\ln|\ln u - 1| = \ln|Cx|, \quad C \neq 0,$$

$$\ln u - 1 = Cx, \quad C \neq 0,$$

$$\ln \frac{y}{x} = Cx + 1, \quad C \neq 0,$$

$$y = xe^{Cx+1}, \quad C \neq 0.$$

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) $$y' = f\left(\frac{ax+by+c}{Ax+By+C}\right).$$

If the system of linear equations

(2) $$ax + by + c = 0, \quad Ax + By + C = 0$$

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

$$\frac{dv}{du} = f\left(\frac{au+bv}{Au+Bv}\right).$$

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

$$\frac{dv}{du} = -\frac{4u+3v}{3u+2v},$$
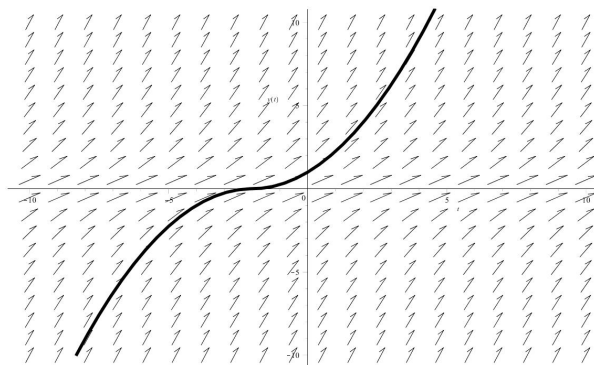
Nevertheless, the existence of a solution is guaranteed by the following theorem, known as *Peano existence theorem*:

Theorem. *Consider a function $f(t, y) : \mathbb{R}^2 \to \mathbb{R}$ 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

$$t_k = t_0 + kh, \quad k = 1, 2, \ldots.$$

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.,

$$\tilde{y}_h(t) = y_{(k)}(t) \quad \text{if } t \in [kh, (k+1)h].$$

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 δ 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 functions on a sufficiently small interval. By the Arzelà-Askoli theorem (see paragraph 7.3.15 on page 499),

which can be solved by further substitution $z = v/u$. We thus obtain

$$z'u + z = -\frac{4 + 3z}{3 + 2z},$$

$$\frac{dz}{du}u = -\frac{2z^2 + 6z + 4}{3 + 2z},$$

$$\frac{2z + 3}{2z^2 + 6z + 4}dz = -\frac{du}{u}$$

provided $z^2 + 3z + 2 \neq 0$. Integrating, we get

$$\frac{1}{2}\ln\left| z^2 + 3z + 2 \right| = -\ln|u| + \ln|C|, \quad C \neq 0,$$

$$\frac{1}{2}\ln\left| \left(z^2 + 3z + 2\right)u^2 \right| = \ln|C|, \quad C \neq 0,$$

$$\ln\left| \left(z^2 + 3z + 2\right)u^2 \right| = \ln C^2, \quad C \neq 0,$$

$$\left(z^2 + 3z + 2\right)u^2 = \pm C^2, \quad C \neq 0.$$

We thus have

$$\left(z^2 + 3z + 2\right)u^2 = D, \quad D \neq 0$$

and returning to the original variables,

$$\left(\frac{v^2}{u^2} + 3\frac{v}{u} + 2\right)u^2 = D, \quad D \neq 0,$$

$$v^2 + 3vu + 2u^2 = D, \quad D \neq 0,$$

$$(y - 1)^2 + 3(y - 1)(x + 1) + 2(x + 1)^2 = D, \quad D \neq 0.$$

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

$$(x + y)(2x + y + 1) = D, \quad D \in \mathbb{R}.$$

$\square$

8.J.8. Find the general solution of the differential equation

$$\left(x^2 + y^2\right)dx - 2xy\,dy = 0.$$

Solution. For $y \neq 0$, simple rearrangements lead to

$$y' = \frac{x^2 + y^2}{2xy} = \frac{1 + \left(\frac{y}{x}\right)^2}{2\frac{y}{x}}.$$

Using the substitution $u = y/x$, we get to the equation

$$u'x + u = \frac{1 + u^2}{2u}.$$

the constructed continuous functions $\tilde{y}_h$ are all in a compact set of functions. So there exists a sequence of values $h_n \to 0$ 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. $\hat{y}_n \to y$ 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 $t_0 + kh_n \leq t$, while y_j and t_j are the points from the definition of the function $\tilde{y}_{h_n}$)

$$\hat{y}_n(t) = y_0 + \sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} f(t_j, y_j)ds + \int_{t_k}^{t} f(t_k, y_k)\,ds.$$

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 $\lim_{n \to \infty} y_n(t)$, thereby obtaining

$$y(t) = \lim_{n \to \infty}\left(y_0 + \int_{t_0}^{t} f(s, \hat{y}_n(s))\,ds\right)$$

$$= y_0 + \int_{t_0}^{t}\left(\lim_{n \to \infty} f(s, \hat{y}_n(s))\right)ds$$

$$= y_0 + \int_{t_0}^{t} f(s, y(s))\,ds,$$

where the uniform convergence $\hat{y}_n(t) \to y(t)$ is employed.

This proves the theorem. $\square$

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 $f, g : \mathbb{R}^3 \to \mathbb{R}$.

A simple example in the plane might be the system of equations

$$x' = -y, \quad y' = x.$$

For $u \neq \pm 1$ and $D = -1/C$, we have

$$\frac{du}{dx}\, x = \frac{1 + u^2 - 2u^2}{2u},$$

$$\frac{2u}{1 - u^2}\, du = \frac{dx}{x},$$

$$-\ln\left|\, 1 - u^2\, \right| = \ln|\, x\, | + \ln|\, C\, |, \quad C \neq 0,$$

$$\ln \frac{1}{|\, 1 - u^2\, |} = \ln|\, Cx\, |, \quad C \neq 0,$$

$$\frac{1}{1 - u^2} = Cx, \quad C \neq 0,$$

$$1 = Cx\left(1 - \frac{y^2}{x^2}\right), \quad C \neq 0,$$

$$-\frac{D}{x} = 1 - \frac{y^2}{x^2}, \quad D \neq 0,$$

$$-Dx = x^2 - y^2, \quad D \neq 0.$$

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

$$y^2 = x^2 + Dx, \quad D \in \mathbb{R}. \qquad \square$$

8.J.9.   Solve

$$y' = x - \tfrac{2y}{x^2 - 1}.$$

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

$$y' \tfrac{x-1}{x+1} + \tfrac{2y}{(x+1)^2} = \tfrac{x(x-1)}{x+1}.$$

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

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 $x : \mathbb{R} \to \mathbb{R}^n$ and a mapping $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$. The validity of the theorem on uniqueness and existence of the solution to such systems can be extended:

EXISTENCE AND UNIQUENESS FOR SYSTEMS OF ODEs

Theorem.   *Consider functions $f_i(t, x_1, \ldots, x_n) : \mathbb{R}^{n+1} \to \mathbb{R}$, $i = 1, \ldots, n$, with continuous partial derivatives. Then, for every point $(t_0, y) \in \mathbb{R}^{n+1}$, $y = (c_1, \ldots, c_n)$, there exists a maximal interval $[t_0 - a, t_0 + b]$, with positive numbers $a, b \in \mathbb{R}$, and a unique function $x(t) : \mathbb{R} \to \mathbb{R}^n$ which is the solution of the system of equations*

$$x_1' = f_1(t, x_1, \ldots, x_n)$$

$$\vdots$$

$$x_n' = f_n(t, x_1, \ldots, x_n)$$

with the initial condition $x(t_0) = y$, i.e.

$$x_1(t_0) = c_1, \ldots, x_n(t_0) = c_n.$$

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 $x(t) = (x_1(t), \ldots, x_n(t))$ 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

$$\|f(t, z_1, \ldots, z_n) - f(t, y_1, \ldots, y_n)\|$$

is bounded from above by the sum

$$\|f(t, z_1, \ldots, z_n) - f(t, y_1, z_2 \ldots, z_n)\| + \ldots$$
$$+ \|f(t, y_1, \ldots, y_{n-1}, z_n) - f(t, y_1, \ldots, y_n)\|.$$

It is recommended to go through the proof of Theorem 8.3.6 from this point of view and to think out the details. $\qquad \square$

$y\frac{x-1}{x+1} = \int \frac{x(x-1)}{x+1}\, dx = \frac{x^2}{2} - 2x + 2\ln|x+1| + C, \quad C \in$
$\mathbb{R}.$

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

$$\frac{dy}{dx} = -\frac{2y}{x^2-1},$$

$$\frac{dy}{y} = -\frac{2}{x^2-1}\, dx,$$

$$\ln|y| = -\ln|x-1| + \ln|x+1| + \ln|C|, \quad C \neq 0,$$

$$\ln|y| = \ln\left|C\frac{x+1}{x-1}\right|, \quad C \neq 0,$$

$$y = C\frac{x+1}{x-1}, \quad C \neq 0,$$

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

$$\frac{C'(x)\,(x+1)(x-1) + C(x)\,(x-1) - C(x)\,(x+1)}{(x-1)^2} = x - \frac{2\,C(x)\,(x+1)}{(x-1)(x^2-1)}.$$

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:

$$y = C(x)\frac{x+1}{x-1} =$$
$$\frac{x+1}{x-1}\left(\frac{x^2}{2} - 2x + 2\ln|x+1| + C\right), \quad C \in \mathbb{R}.$$

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.

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 α) 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 γ). Further, consider an interaction of the predator and the prey which is expected to be proportional to the number of both with a certain coefficient β, 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 pray, y the predator, with positive constants α, β, γ, δ

$$x' = \alpha x - \beta y x$$
$$y' = -\gamma y + \delta \beta x y.$$

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 $(x_0, y_0) = (1, 1.5)$
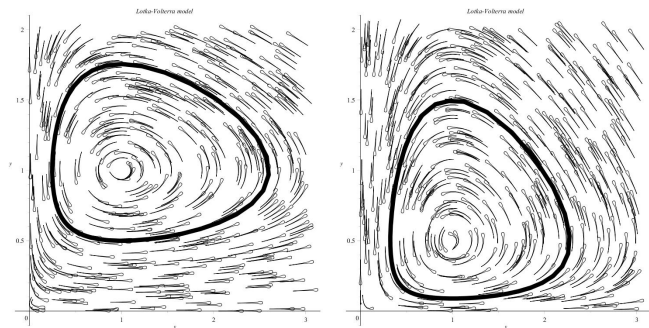
In both cases, the system is quite stable in the vicinity of the initial condition, and it would be very stable for $(x_0, y_0) = (1, 1)$ or $(1, 0.5)$, respectively. But their developement differs in speed — the depicted solution cycles close in the times about $t = 12$ in the first case and $t = 5$ in the other one.

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

$$y = C e^{\int a(x)\, dx}, \quad C \in \mathbb{R}.$$

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 $C \in \mathbb{R}$. □

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

$$\left(y\, e^{x^2} \right)' = \cos x,$$

$$y\, e^{x^2} = \int \cos x\, dx,$$

$$y\, e^{x^2} = \sin x + C, \quad C \in \mathbb{R},$$

$$y = e^{-x^2} \left(\sin x + C \right), \quad C \in \mathbb{R}.$$

□

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

$$u' = (1 - r)\left[a(x)u + b(x) \right].$$

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

$$u = e^{-\ln|x|} \left[\int -3x\, e^{\ln|x|}\, dx \right],$$

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.

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 the to L_∞ 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.[7]

Consider two systems of equations written in the vector form

$$(1) \qquad x' = f(t, x), \quad y' = g(t, y)$$

and assume that the mappings $f, g : U \subset \mathbb{R}^{n+1} \to \mathbb{R}^n$ 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

$$C = \sup_{x \neq y;\ (t,x),\ (t,y) \in U} \frac{|f(t,x) - g(t,y)|}{|x - y|}$$

$$B = \sup_{(t,x) \in U} |f(t,x) - g(t,x)|$$

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,*

$$|x(t) - y(t)| \le |x_0 - y_0|\, e^{C|t - t_0|} + \frac{B}{C} \left(e^{C|t - t_0|} - 1 \right).$$

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

$$|x(t) - y(t)| \le |x_0 - y_0| + \int_0^t |f(s, x(s)) - g(s, y(s))|\, ds.$$

---

[7]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.

CHAPTER 8. CALCULUS WITH MORE VARIABLES

where $\ln|x|$ was obtained as an (arbitrary) antiderivative to $1/x$. Furhter,

$$u = e^{\ln\frac{1}{|x|}}\left[\int -3x\,e^{\ln|x|}\,dx\right],$$
$$u = \frac{1}{|x|}\left[\int -3x\,|x|\,dx\right].$$

The absolute value can be replaced with a sign that can be canceled, i. e., it suffices to consider

$$u = \tfrac{1}{x}\left[\int -3x^2\,dx\right] = \tfrac{1}{x}\left[-x^3 + C\right], \quad C \in \mathbb{R}.$$

Returning to the original variable, we get

$$y = \tfrac{1}{u} = \tfrac{x}{C-x^3}, \quad C \in \mathbb{R}.$$

The excluded case $y \equiv 0$ is a singular solution (which, of course, is true for every Bernoulli equation with r positive).
□

8.J.12. Interchanging the variables, solve the equation

$$y\,dx - \left(x + y^2\sin y\right)\,dy = 0.$$

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:

$$y' = \tfrac{y}{x+y^2\sin y}.$$

This equation is not of any of the previous types, so we rewrite it as follows:

$$\frac{dy}{dx} = \frac{y}{x + y^2\sin y},$$
$$\frac{dx}{dy} = \left(\frac{y}{x+y^2\sin y}\right)^{-1} = \frac{x}{y} + y\sin y,$$
$$x' = \frac{1}{y}x + y\sin y.$$

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 .

The integrand can be further estimated as follows:

$$|f(s, x(s)) - g(s, y(s))| \le$$
$$\le |f(s, x(s)) - f(s, y(s))| + |f(s, y(s)) - g(s, y(s))|$$
$$\le C\,|x(s) - y(s)| + B$$

If $F(t) = |x(t) - y(t)|$, $\alpha = |x_0 - y_0|$, then

$$F(t) \le \alpha + \int_0^t (C\,F(s) + B)\,ds.$$

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 $[0, t_{max}]$*

$$F(t) \le \alpha(t) + \int_0^t \beta(s)F(s)\,ds$$

for some real-valued functions $\alpha(t)$, $\beta(t)$, with $\beta(t) \ge 0$. Then

$$F(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s)\,e^{\int_s^t \beta(r)\,dr}\,ds$$

for all $t \in [0, t_{max}]$. Moreover, if additionally $\alpha(t)$ is non-decreasing, then

$$F(t) \le \alpha(t)\,e^{\int_0^t \beta(s)\,ds}.$$

PROOF OF THE LEMMA. Write

$$G(t) = e^{-\int_0^t \beta(s)\,ds}.$$

By the first assumption of the theorem,

$$\frac{d}{dt}\left(G(t)\int_0^t \beta(s)F(s)\,ds\right) =$$
$$= \beta(t)G(t)\left(F(t) - \int_0^t \beta(s)F(s)\,ds\right)$$
$$\le \alpha(t)\beta(t)G(t).$$

Integrating with respect to t and dividing by the non-zero function $G(t)$ gives

$$\int_0^t \beta(s)F(s)\,ds \le \int_0^t \alpha(s)\beta(s)\frac{G(s)}{G(t)}\,ds,$$

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:

$$F(t) \le \alpha(t)\Big(1 + \int_0^t \beta(s)\,e^{\int_s^t \beta(r)\,dr}\,ds\Big).$$

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

$$F(t) \le \alpha(t)\Big(1 - \int_0^t \frac{d}{ds}\,e^{\int_s^t \beta(r)\,dr}\,ds\Big)$$
$$= \alpha(t)\Big(1 + e^{\int_s^t \beta(r)\,dr} - 1\Big),$$

K. Practical problems leading to differential equations

8.K.1. A water purification plant with volume $2000\,\mathrm{m}^3$ was contaminated with lead which is spread in the water with density $10\,\mathrm{g/m}^3$. Water is flowing in and out of the basin at $2\,\mathrm{m}^3/\mathrm{s}$. In what time does the amount of lead in the basin decrease below $10\,\mu\mathrm{g/m}^3$ (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 (m^3), the speed of the water's flow by v (m^3/s). In an infinitesimal (infinitely small) time unit $\mathrm{d}t$, $\frac{m}{V}\cdot v\,\mathrm{d}t$ grams of lead runs out of the basin, so we can construct the differential equation

$$\mathrm{d}m = -\frac{m}{V}\cdot v\,\mathrm{d}t$$

for the change of the lead's mass in the basin. Separating the variables, we get the equation

$$\frac{\mathrm{d}m}{m} = -\frac{v}{V}\,\mathrm{d}t.$$

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 $t \doteq 6\,\mathrm{h}\ 35\,\mathrm{min}$. $\square$

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. $\square$

and the second proposition of the lemma is also proved. $\square$

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

$$\tilde{F}(t) \leq \frac{B}{C} + \alpha + \int_0^t C\tilde{F}(s)\,ds.$$

This is the assumption of Gronwall's inequality with even constant parameters, so by the second claim of the lemma,

$$F(t) + \frac{B}{C} \leq \left(\alpha + \frac{B}{C}\right)e^{\int_0^t C\,ds},$$

which is the statement

$$F(t) \leq \alpha\,e^{Ct} + \frac{B}{C}(e^{Ct} - 1)$$

as desired. $\square$

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 $y_1, \ldots, y_n$. Then there is the same system of equations for the curve $\tilde{y}'(t) = (y_0(t), y_1(t), \ldots, y_n(t))$ as

$$y_0' = 1$$
$$y_1' = f_1(y_0, y_1, \ldots, y_n)$$
$$\vdots$$
$$y_n' = f_n(y_0, y_1, \ldots, y_n)$$

with the initial conditions

$$y_0(t_0) = t_0,\ y_1(t_0) = x_1, \ldots, y_n(t_0) = x_n.$$

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 λ and with initial conditions

$$(1) \qquad\qquad y' = f(y, \lambda),\ y(t_0) = x.$$

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

$$f(t) = \frac{K}{1 + Le^{-Kkt}}.$$

$\square$

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 $^{239}_{94}Pu$ 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. $\square$

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 ms^{-1}, the resistance constant is $k = 0.5 \text{ s}^{-1}$. What will the speed of the object be in three seconds?

Solution.

$$v = \frac{g}{k} - \left(\frac{g}{k} - v_0\right) e^{-kt},$$

$v(3) = 20 - 15e^{-\frac{3}{2}}$ ms^{-1} after substitution. $\square$

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 λ), 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 the Jacobi matrix $D^1 f(y)$ of the mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ 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:

$$y^{(3)} = D^2 f(y)\big(f(y), f(y)\big) + \big(D^1 f(y)\big)^2 \cdot f(y).$$

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 λ 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:

$$\begin{aligned}
\Phi'(t, x) &= \frac{d}{dt}\big(D_x^1 y(t, x)\big) = D_x^1\big(y'(t, x)\big) \\
&= D^1 f(y(t, x)) \cdot D_x^1 y(t, x) \\
&= D^1 f(y(t, x)) \cdot \Phi(t, x).
\end{aligned}$$

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

$$(2) \qquad \Phi'(t, x) = F(t, x) \cdot \Phi(t, x), \quad \Phi(0, x) = E,$$

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.

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. $\qquad\square$

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. $\qquad\bigcirc$

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 $t \geq 0$? $\qquad\bigcirc$

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. $\qquad\bigcirc$

Further practical problems that lead to differential equations can be found on page 595.

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 case $k = 1$, we get the so-called oscillation equation

$$y''(t) = -y(t).$$

This equation corresponds to the system of equations

$$x'(t) = -y(t), \quad y'(t) = x(t)$$

Theorem. *Consider an open subset $U \subset \mathbb{R}^{n+k}$ and a mapping $f : U \to \mathbb{R}^n$ 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.

PROOF. Consider a general system dependent on parameters, but viewed as an ordinary autonomous system with no parameters. More explicitely, 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) wit 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

$$\lim_{h \to 0} \frac{1}{\|h\|} G(t, h) = 0.$$

Several estimates are needed for this purpose.

First, from the latter theorem 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

$$\frac{R(y, z)}{\|y - z\|} \to 0 \text{ for } \|y - z\| \to 0.$$

This implies the crucial estimate. In the first equality substitute in the expression of solutions in terms of fixed

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 τ, 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 α 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.
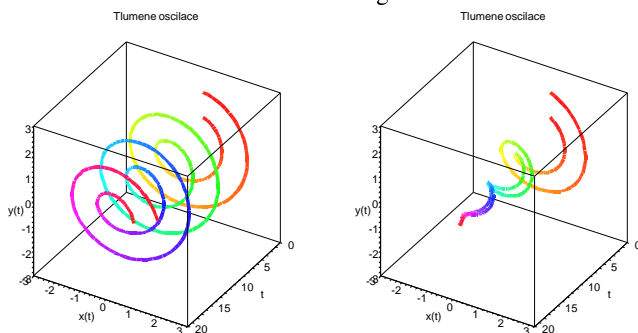
The oscillations are expressed by the y-axis values; the x-axis values describe the speed of the motion.

8.L.2. Undamped oscillation. Find the function $y(t)$ which satisfies the following differential equation and initial conditions:

$$y''(t) + 4y(t) = f(t), \ y(0) = 0, \ y'(0) = -1,$$

where the function $f(t)$ is piecewise continuous:

$$f(t) = \begin{cases} \cos(2t) & \text{for } 0 \le t < \pi, \\ 0 & \text{for } t \ge \pi. \end{cases}$$

Solution. This problem is a model of undamped oscillation of a spring (omitting friction, non-linearities in the toughness

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 summand).

$$G(t, h) = \left\| x_0 + h + \int_0^t f(y(s, x_0 + h))ds - x_0 \right.$$
$$\left. - \int_0^t f(y(s, x_0))ds - \Phi(t, x_0)(h) \right\|$$
$$= \left\| \int_0^t \Big(f(y(s, x_0 + h)) - f(y(s, x_0)) \right.$$
$$\left. - F(s, x_0)\Phi(s, x_0)(h) \Big)ds \right\|$$
$$\le \int_0^t \| f(y(s, x_0 + h)) - f(y(s, x_0))$$
$$- F(s, x_0)\Phi(s, x_0)(h)\| \, ds$$
$$\le \int_0^t \|F(s, x_0)\| \, \|y(s, x_0+h) - y(s, x_0) - \Phi(s, x_0)(h)\| \, ds$$
$$+ \int_0^t \|R(y(s, x_0 + h), y(s, x_0))\| \, ds,$$

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

$$\|F(t, x_0)\| \le B,$$

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

$$\|R(y(t, x_0 + h), y(t, x_0))\| \le \varepsilon \|y(t, x_0 + h) - y(t, x_0)\|$$
$$\le \|h\|\varepsilon\, \mathrm{e}^{CT}.$$

Therefore, the estimate on $G(t, h)$ can be improved as follows:

$$G(t, h) \le B \int_0^t G(s, h) \, ds + \varepsilon \|h\| \, \mathrm{e}^{CT}.$$

Gronwall's lemma now gives

$$G(t, h) \le \varepsilon \|h\| \, \mathrm{e}^{(C+B)T}.$$

This implies that $\lim_{h \to 0} \frac{1}{\|h\|} G(t, h) = 0$ as requested. $\qquad \square$

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

$$(1) \qquad\qquad y' = f(y), \quad y(t_0) = y_0$$

of the spring, and other factors) which is initiated by an outer force.

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

$$y(t) = -\tfrac{1}{2}\sin(2t) + \tfrac{1}{4}t\sin(2t) + \mathcal{L}^{-1}\left(e^{-\pi s}\frac{s}{(s^2+4)^2}\right).$$

However, by formula (1), we have

$$\begin{aligned}
\mathcal{L}^{-1}\left(e^{-\pi s}\frac{s}{(s^2+4)^2}\right) &= \tfrac{1}{4}\mathcal{L}^{-1}(e^{-\pi s}\mathcal{L}(t\sin(2t))) \\
&= (t-\pi)\sin(2(t-\pi)) \cdot H_\pi(t).
\end{aligned}$$

Since Heaviside's function is zero for $t < \pi$ and equal to 1 for $t > \pi$, we get the solution in the form

$$y(t) = \begin{cases} -\tfrac{1}{2}\sin(2t) + \tfrac{1}{4}t\sin(2t) & \text{for } 0 \le t < \pi \\ \frac{\pi-2}{4}\sin(2t) & \text{for } t \ge \pi \end{cases}$$

$\square$

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

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_{n-1}y' + a_n y = f(x)$$

for certain constants $a_1, \ldots, a_n \in \mathbb{R}$. 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

$$\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n.$$

Each real root λ with multiplicity k corresponds to the k solutions

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.

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 dependancy on the initial condition.*

PROOF. The idea of the proof is identical as in the simple one-dimensional case in 6.2.15. 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

$$y^{(k)}(0) = P_k(f(y(0)), \ldots, \partial_\beta f(y(0)), \ldots)$$

for these multivariate vector valued polynomials (the multi-indices β 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 dependancy 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 $\partial_\alpha g(0) \ge |\partial_\alpha f(0)|$, for all multi-indices α. 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:

$$\begin{aligned}
z^{(k)}(0) &= P_k(g(0), \ldots, \partial_\beta g(0), \ldots) \\
&\ge P_k(|f(0)|, \ldots, |\partial_\beta f(0)|) \ge |y^{(k)}(0)|.
\end{aligned}$$

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 $|\frac{1}{\alpha!}\partial_\alpha f_i(0)r^{|\alpha|}| \le C$, for all $i = 1, \ldots, n$ and mutli-indices α. This means $|\partial_\alpha f_i(0)| \le C\frac{\alpha!}{r^{|\alpha|}}$. In the 1-dimensional case, we considered the multiple of a geometric series $g(z) = C\frac{r}{r-z}$ with the right

$$\mathrm{e}^{\lambda x}, x\,\mathrm{e}^{\lambda x}, \ldots, x^{k-1}\,\mathrm{e}^{\lambda x}$$

and every pair of complex roots $\lambda = \alpha \pm \mathrm{i}\beta$ with multiplicity k corresponds to the k pairs of solutions

$$\mathrm{e}^{\alpha x}\cos\left(\beta x\right), x\,\mathrm{e}^{\alpha x}\cos\left(\beta x\right), \ldots, x^{k-1}\,\mathrm{e}^{\alpha x}\cos\left(\beta x\right),$$
$$\mathrm{e}^{\alpha x}\sin\left(\beta x\right), x\,\mathrm{e}^{\alpha x}\sin\left(\beta x\right), \ldots, x^{k-1}\,\mathrm{e}^{\alpha x}\sin\left(\beta x\right).$$

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:

$$y = C_1\mathrm{e}^{4x} + C_2 x\,\mathrm{e}^{4x} + C_3\mathrm{e}^{-3x}, \quad C_1, C_2, C_3 \in \mathbb{R}. \;\square$$

8.L.4. Compute

$$y''' + y'' + 9y' + 9y = \mathrm{e}^x + 10\cos\left(3x\right).$$

Solution. First, we will solve the corresponding homogeneous equation. The characteristic polynomial is equal to

$$\lambda^3 + \lambda^2 + 9\lambda + 9,$$

with roots $\lambda_1 = -1$, $\lambda_2 = 3\mathrm{i}$, $\lambda_3 = -3\mathrm{i}$. The general solution of the corresponding homogeneous equation is thus

$$y = C_1\mathrm{e}^{-x} + C_2\cos\left(3x\right) + C_3\sin\left(3x\right), \quad C_1, C_2, C_3 \in \mathbb{R}.$$

The solution of the non-homogeneous equation is of the form

$$y =$$
$$C_1\mathrm{e}^{-x} + C_2\cos\left(3x\right) + C_3\sin\left(3x\right) + y_p, \quad C_1, C_2, C_3 \in \mathbb{R}$$

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

$$P_n(x)\,\mathrm{e}^{\alpha x},$$

where P_n is a polynomial of degree n, then there is a particular solution of the form

$$y_p = x^k\,R_n(x)\,\mathrm{e}^{\alpha x},$$

where k is the multiplicity of α 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

$$\mathrm{e}^{\alpha x}\left[P_m(x)\cos\left(\beta x\right) + S_n(x)\sin\left(\beta x\right)\right],$$

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

$$y_p = x^k\,\mathrm{e}^{\alpha x}\left[R_l(x)\cos\left(\beta x\right) + T_l(x)\sin\left(\beta x\right)\right],$$

derivatives $g^{(n)} = C\frac{n!}{r^n}$. Now the most similar mapping is $g(z_1, \ldots, z_n) = (g_1(z_1, \ldots, z_n), \ldots, g_n(z_1, \ldots, z_n))$ with all the components g_i equal to

$$h : \mathbb{R}^n \to \mathbb{R}, \quad h(z_1, \ldots, z_n) = C\frac{r}{r - z_1 - \cdots - z_n}.$$

Then the values of all the partial derivatives with $|\alpha| = k$ at $z = 0$ are

$$\partial_\alpha h(0) = Crk!(r - z_1 - \cdots - z_n)^{-k-1}{}_{|z=0} = C\frac{k!}{r^k},$$

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 compoments 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,

$$u'(t) = h(u(t), u(t), \ldots, u(t)) = C\frac{r}{r - nu(t)}.$$

This is nearly exactly the same equation as the one in 6.2.15 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.
$$\square$$

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 $X : U \to TU$ which is a section of the projection p, i.e., $p \circ X = \mathrm{id}_U$. The *derivative in the direction of the vector field X* is defined for all differentiable functions g on U by

$$X(g) : U \to \mathbb{R}, \qquad X(g)(x) = d_{X(x)}g = dg(x)(X(x)).$$

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:

$$(1) \qquad\qquad X(gh) = hX(g) + gX(h).$$

In fixed coordinates, $X(x) = (X_1(x), \ldots X_n(x))$ and

$$X(g)(x) = X_1(x)\frac{\partial g}{\partial x_1}(x) + \cdots + X_n(x)\frac{\partial g}{\partial x_n}(x).$$

where k is the multiplicity of $\alpha + \mathrm{i}\beta$ 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 $y = \mathrm{e}^x$, which has particular solution $y_{p_1}(x) = A\mathrm{e}^x$ for some $A \in \mathbb{R}$. Since

$$y_{p_1}(x) = y'_{p_1}(x) = y''_{p_1}(x) = y'''_{p_1}(x) = A\mathrm{e}^x,$$

substitution into the original equation, whose right-hand side contains only the function $y = \mathrm{e}^x$, leads to

$$20A\mathrm{e}^x = \mathrm{e}^x, \quad \text{i. e.} \quad A = \tfrac{1}{20}.$$

For the right-hand side with the function $y = 10\cos(3x)$, we are looking for a particular solution in the form

$$y_{p_2}(x) = x\left[B\cos(3x) + C\sin(3x)\right].$$

Recall that the number $\lambda = 3\mathrm{i}$ was obtained as a root of the characteristic polynomial. We can easily compute the derivatives

$$y'_{p_2}(x) = [B\cos(3x) + C\sin(3x)]$$
$$+ x\left[-3B\sin(3x) + 3C\cos(3x)\right],$$
$$y''_{p_2}(x) = 2\left[-3B\sin(3x) + 3C\cos(3x)\right]$$
$$+ x\left[-9B\cos(3x) - 9C\sin(3x)\right],$$
$$y'''_{p_2}(x) = 3\left[-9B\cos(3x) - 9C\sin(3x)\right]$$
$$+ x\left[27B\sin(3x) - 27C\cos(3x)\right].$$

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.,

$$y_{p_2}(x) = x\left[-\tfrac{1}{2}\cos(3x) + \tfrac{1}{6}\sin(3x)\right].$$

Altogether, the general solution is

$$y = C_1\mathrm{e}^{-x} + C_2\cos(3x) + C_3\sin(3x) + \frac{1}{20}\mathrm{e}^x$$
$$-\frac{1}{2}x\cos(3x) + \frac{1}{6}x\sin(3x), \quad C_1, C_2, C_3 \in \mathbb{R}.$$

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

$$\mathrm{Fl}_t^X : \mathbb{R}^n \to \mathbb{R}^n,$$

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

$$\mathrm{Fl}_{t+s}^X(x) = \mathrm{Fl}_t^X \circ \mathrm{Fl}_s^X(x),$$

whenever both sides exist. In particular, the mappings Fl_s^X and Fl_t^X always commute.

Moreover, the mapping $\mathrm{Fl}_t^X(x)$ 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 $\mathrm{Fl}_t^X : \mathbb{R}^n \to \mathbb{R}^n$ with inverse diffeomorphisms Fl_{-t}^X.

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 $y_0 = axe^{-2x}$ for $a \in \mathbb{R}$.

Substituting into the original equation, we obtain

$$a[-4e^{-2x} + 4xe^{-2x} + 3(e^{-2x} - 2xe^{-2x}) + 2xe^{-2x}] = e^{-2x},$$

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 $c_1 e^{-x} + c_2 e^{-2x} - xe^{-2x}$, $c_1, c_2 \in \mathbb{R}$. $\qquad\square$

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 $c_1 + c_2 e^{-x}$, 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 $c_1 + c_2 e^{-x} + x$, $c_1, c_2 \in \mathbb{R}$. $\qquad\square$

A simple example of a complete vector field is the field $X(x) = \frac{\partial}{\partial x_1}$. Its flow is given by

$$\mathrm{Fl}_t^X(x_1, \ldots, x_n) = (x_1 + t, x_2, \ldots, x_n).$$

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 $X : U \subset \mathbb{R}^n \to \mathbb{R}^n$ where $X(x_0) = 0$ are called *singular points* of the vector field X. Clearly $\mathrm{Fl}_t^X(x_0) = x_0$ 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 $F : \mathbb{R}^n \to \mathbb{R}^n$ (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

$$y'(t) = \big(F(x(t))\big)'(t) = D^1 F(x(t)) \cdot x'(t)$$
$$= D^1 F(x(t)) \cdot X(x(t)).$$

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:

$$\mathrm{Fl}_t^Y \circ F(x) = F \circ \mathrm{Fl}_t^X(x).$$

By fixing $x = x_0$ and writing $x(t) = \mathrm{Fl}_t^X(x_0)$, 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.

<div align="center">THE FLOWBOX THEOREM</div>

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

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 $c_1 e^{-2x} + c_2 e^{-3x}$, $c_1, c_2 \in \mathbb{R}$. 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

$$c_1 e^{-2x} + c_2 e^{-3x} + xe^{-2x}.$$

$\square$

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 $c_1 + c_2 e^x$, 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

$$c_1 + c_2 e^x - 5x.$$

$\square$

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

$$y = C_1(x)\, y_1(x) + C_2(x)\, y_2(x) + \cdots + C_n(x)\, y_n(x),$$

where $y_1, \ldots, y_n$ give the general solution of the corresponding homogeneous equation and the functions $C_1(x), \ldots, C_n(x)$ can be obtained from the system

$$C_1'(x)\, y_1(x) + \cdots + C_n'(x)\, y_n(x) = 0,$$
$$C_1'(x)\, y_1'(x) + \cdots + C_n'(x)\, y_n'(x) = 0,$$
$$\vdots$$
$$C_1'(x)\, y_1^{(n-2)}(x) + \cdots + C_n'(x)\, y_n^{(n-2)}(x) = 0,$$
$$C_1'(x)\, y_1^{(n-1)}(x) + \cdots + C_n'(x)\, y_n^{(n-1)}(x) = f(x).$$

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 $X(0) = \frac{\partial}{\partial x_1}(0)$.

With such coordinates, write the flow of the field X going through the point $(x_1, \ldots, x_n)$ at time $t = 0$ as $x_i(t) = \varphi_i(t, x_1, \ldots, x_n)$, $i = 1, \ldots, n$. Next, define the components of $F = (f_1, \ldots, f_n)$ as

$$f_i(x_1, \ldots, x_n) = \varphi_i(x_1, 0, x_2, \ldots, x_n).$$

This follows the strategy. Since $X(0, \ldots, 0) = \frac{\partial}{\partial x_1}$

$$\frac{\partial F}{\partial x_1}(0, \ldots, 0) = \frac{d}{dt}_{|0} (\varphi_1(t, 0, \ldots, 0), \ldots \varphi_n(t, 0, \ldots, 0))$$
$$= (1, 0, \ldots, 0),$$

while the flow Fl_0^X at the time $t = 0$ yields

$$\varphi_i(0, 0, x_2, \ldots, x_n) = (0, x_2, \ldots, x_n),$$

and in particular

$$\frac{\partial F}{\partial x_i}(0, \ldots, 0) = (0, \ldots, 1, \ldots, 0), \quad i = 2, \ldots, n.$$

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 8.1.22).

Directly from the definition of the mapping F in terms of the flow of the vector field X, the flow of the transformed field Y is expressed in the new coordinates $(y_1, \ldots, y_n)$ as

$$\mathrm{Fl}_t^Y(y_1, \ldots, y_n) = (y_1 + t, y_2, \ldots, y_n).$$

Verify this by yourselves in detail! $\square$

8.3.16. Higher-order equations. An ordinary differential equation of order k (solved with respect to the highest derivative) is an equation

$$(1) \qquad y^{(k)} = f(t, y, y', \ldots, y^{(k-1)}),$$

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:

$$y_0(t) = y(t), \ y_1(t) = y_0'(t), \ldots, \ y_{k-1}(t) = y_{k-2}'(t).$$

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 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

$$C_1'(x)\, e^x + C_2'(x)\, x\, e^x = 0,$$

$$C_1'(x)\, e^x + C_2'(x)\, [e^x + x\, e^x] = \frac{e^x}{x^2 + 1}.$$

We can compute the unknowns $C_1'(x)$ and $C_2'(x)$ using Cramer's rule. It follows from

$$\begin{vmatrix} e^x & x\, e^x \\ e^x & e^x + x\, e^x \end{vmatrix} = e^{2x},$$

$$\begin{vmatrix} 0 & x\, e^x \\ \frac{e^x}{x^2+1} & e^x + x\, e^x \end{vmatrix} = -x\,\frac{e^{2x}}{x^2 + 1},$$

$$\begin{vmatrix} e^x & 0 \\ e^x & \frac{e^x}{x^2+1} \end{vmatrix} = \frac{e^{2x}}{x^2 + 1}$$

that

$$C_1(x) = -\int \frac{x}{x^2 + 1}\, dx = -\frac{1}{2}\ln\left(x^2 + 1\right) + C_1, \quad C_1 \in \mathbb{R},$$

$$C_2(x) = \int \frac{dx}{x^2 + 1} = \arctan x + C_2, \quad C_2 \in \mathbb{R}.$$

Hence, the general solution is

$$y = C_1 e^x + C_2 x\, e^x - \tfrac{1}{2}\, e^x \ln\left(x^2 + 1\right) + x\, e^x \arctan x, \quad C_1, C_2 \in \mathbb{R}.$$

$\square$

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

$$c_1 e^{2x} + c_2 e^{-x} + c_3 x e^{-x} - \frac{1}{2}e^x.$$

Substituting into the original conditions, we get the only satisfactory function,

$$\frac{2}{9}e^{2x} + \frac{5}{18}e^{-x} + \frac{1}{3}x e^{-x} - \frac{1}{2}e^x.$$

$\square$

Further problems concerning higher-order differential equations can be found on page

the system of equations

$$y_0' = y_1$$
$$y_1' = y_2$$
$$\vdots$$
$$y_{k-2}' = y_{k-1}$$
$$y_{k-1}' = f(t, y_0, y_1, \ldots, y_{k-1}).$$

Hence the following direct corollary of the theorems :

SOLUTIONS OF HIGHER-ORDER ODEs

Theorem. *Consider a function $f(t, y_0, \ldots, y_{k-1}) : U \subset \mathbb{R}^{k+1} \to \mathbb{R}$ with continuous partial derivatives on an open set U. Then for every point $(t_0, z_0, \ldots, z_{k-1}) \in U$, 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 $y(t) : I_{max} \to \mathbb{R}$ which is a solution of the k-th order equation*

$$y^{(k)} = f(t, y, y', \ldots, y^{(k-1)})$$

with the initial condition

$$y(t_0) = z_0, y'(t_0) = z_1, \ldots, y^{(k-1)}(t_0) = z_{k-1}.$$

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 ℓ 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* $y(t) \mapsto D(y)(t)$:

$$D(y)(t) = a_k(t)y^{(k)}(t) + \cdots + a_1(t)y'(t) + a_0(t)y(t).$$

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,

$$D(y_1 + y_2)(t) = D(y_1)(t) + D(y_2)(t).$$

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.

$$(1) \qquad \mathcal{L}(y')(s) = s\mathcal{L}(y)(s) - y(0)$$

$$\mathcal{L}(y'')(s) = s^2\mathcal{L}(y) - sy(0) - y'(0)$$

and, by induction:

$$\mathcal{L}(y^{(n)})(s) \quad = \quad s^n\mathcal{L}(y)(s) \quad - \\ \sum_{i=1}^{n} s^{n-i} y^{(i-1)}(0) \,. \qquad \square$$

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.H.8 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

$$y(t) = \tfrac{1}{8}\sin 2t - \tfrac{1}{4}t\cos 2t \,. \qquad \square$$

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

$$\frac{50}{(s^2 + 1)(s + 3)^2} = \frac{As + B}{s^2 + 1} + \frac{C}{s + 3} + \frac{D}{(s + 3)^2},$$

so

$$50 = (As + B)(s + 3)^2 + C(s^2 + 1)(s + 3) + D(s^2 + 1).$$

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 λ can satisfy an equation of order k. Substitution yields

$$D(e^{\lambda t}) = \left(a_k\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0\right)e^{\lambda t}\,.$$

The parameter λ leads to a solution of a linear differential equation with constant coefficients if and only if λ is a root of the *characteristic polynomial* $a_k\lambda^k + \cdots + a_1\lambda + a_0$.

If this polynomial has k distinct roots, then we have the basis of the whole vector space of solutions. Otherwise, if λ is a multiple root, then direct calculation, making use of the fact that λ is then a root of the derivative of the characteristic polynomial as well, yields that the function $y(t) = t\,e^{\lambda t}$ is

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

$$50 = 9B + 3C + D = 12C + 3C + 5$$

$$\text{hence} \quad C = 3, \ B = 4, \ A = -3.$$

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

$$\mathcal{L}(y)(s) = \frac{-3s + 4}{s^2 + 1} + \frac{3}{s + 3} + \frac{5}{(s + 3)^2} + \frac{1}{s + 3} + \frac{7}{(s + 3)^2}$$

$$= \frac{-3s}{s^2 + 1} + \frac{4}{s^2 + 1} + \frac{4}{s + 3} + \frac{12}{(s + 3)^2}.$$

Now, the inverse Laplace transform yields the solution in the form

$$y(t) = -3\cos t + 4\sin t + 4e^{-3t} + 12te^{-3t}.$$

$\square$

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

$$\mathcal{L}(f'')(s) = s^2 \mathcal{L}(f)(s) - s\lim_{t \to 0+} f(t) - \lim_{t \to 0+} f'(t)$$

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

$$\frac{s}{(s^2+1)(s^2+\pi^2)} = \frac{1}{\pi^2 - 1}\left(\frac{s}{s^2+1} - \frac{s}{s^2+\pi^2}\right).$$

also a solution. Similarly, for higher multiplicities ℓ, There are ℓ distinct solutions $e^{\lambda t}, te^{\lambda t}, \ldots, t^{\ell-1}e^{\lambda t}$.

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 $Y(t)$:

$$(1) \qquad Y'(t) = A \cdot Y(t).$$

Clearly this is a strict analogy to the iterative models in chapter 3.

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 k-th summand in $b_{ij}(t)$ is at most $\frac{t^k}{k!}n^k C^k$ in absolute value. Hence, every series $b_{ij}(t)$ is necessarily absolutely and uniformly convergent, and it is bound above by the value e^{tnC}. 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

$$\frac{d}{dt}\left(e^{tA}\right) = A\,e^{tA}$$

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

The above expression of $\mathcal{L}\left(\cos\left(bt\right)\right)\left(s\right)$ and the already proved formula

$$\mathcal{L}\left(\sin t\right)\left(s\right) = \frac{1}{s^2+1}$$

then yield the wanted solution

$$y(t) = \frac{1}{\pi^2-1}\left(\cos t - \cos\left(\pi t\right)\right) + c_1\cos t + c_2\sin t \,.$$
$\square$

8.M.5. Solve the system of differential equations

$$x''(t) + x'(t) = y(t) - y''(t) + \mathrm{e}^t, \quad x'(t) + 2x(t) = -y(t) + y'(t) + \mathrm{e}^{-t}$$

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

$$\mathcal{L}\left(\mathrm{e}^{\pm t}\right)\left(s\right) = \frac{1}{s\mp 1},$$

transforms the first equation to

$$s^2\mathcal{L}\left(x\right)\left(s\right) - s\lim_{t\to 0+}x(t) - \lim_{t\to 0+}x'(t) + s\mathcal{L}\left(x\right)\left(s\right) - \lim_{t\to 0+}x(t) = =$$
$$\mathcal{L}\left(y\right)\left(s\right) - \left(s^2\mathcal{L}\left(y\right)\left(s\right) - s\lim_{t\to 0+}y(t) - \lim_{t\to 0+}y'(t)\right) + \frac{1}{s-1}$$

and the second one to

$$s\mathcal{L}\left(x\right)\left(s\right) - \lim_{t\to 0+}x(t) + 2\mathcal{L}\left(x\right)\left(s\right) =$$
$$= -\mathcal{L}\left(y\right)\left(s\right) + s\mathcal{L}\left(y\right)\left(s\right) - \lim_{t\to 0+}y(t) + \frac{1}{s+1}.$$

Evaluating the limits (according to the initial conditions), we obtain the linear equations

$$s^2\mathcal{L}\left(x\right)\left(s\right) - 1 + s\mathcal{L}\left(x\right)\left(s\right) = \mathcal{L}\left(y\right)\left(s\right) - s^2\mathcal{L}\left(y\right)\left(s\right) + \frac{1}{s-1}$$

and

$$s\mathcal{L}\left(x\right)\left(s\right) + 2\mathcal{L}\left(x\right)\left(s\right) = -\mathcal{L}\left(y\right)\left(s\right) + s\mathcal{L}\left(y\right)\left(s\right) + \frac{1}{s+1}$$

with the only solution

$$\mathcal{L}\left(x\right)\left(s\right) = \frac{2s-1}{2(s-1)(s+1)^2}, \quad \mathcal{L}\left(y\right)\left(s\right) = \frac{3s}{2(s^2-1)^2}.$$

Once again, we perform partial fraction decomposition, getting

$$\mathcal{L}\left(x\right)\left(s\right) = \frac{1}{8}\frac{1}{s-1} + \frac{3}{4}\frac{1}{(s+1)^2} - \frac{1}{8}\frac{1}{s+1} = \frac{3}{4}\frac{1}{(s+1)^2} + \frac{1}{4}\frac{1}{s^2-1}.$$

Since we have already computed that

$$\mathcal{L}\left(t\,\mathrm{e}^{-t}\right)\left(s\right) = \frac{1}{(s+1)^2}, \quad \mathcal{L}\left(\sinh t\right)\left(s\right) = \frac{1}{s^2-1},$$
$$\mathcal{L}\left(t\sinh t\right)\left(s\right) = \frac{2s}{(s^2-1)^2},$$

we get

$$x(t) = \frac{3}{4}\,t\,\mathrm{e}^{-t} + \frac{1}{4}\sinh t, \quad y(t) = \frac{3}{4}\,t\sinh t.$$

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 $y = \mathrm{e}^t$, $y = \sinh t$ and $y = t\sinh t$ were obtained only for $s > 1$). $\square$

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 $A \in \mathrm{Mat}_n(\mathbb{R})$,

(2) $$y'(t) = A \cdot y(t),$$

for an unknown function $y : \mathbb{R} \to \mathbb{R}^n$, 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 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 Φ 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 Φ as a sum of two commuting linear mappings $\Phi = \Phi_d + \Phi_n$ with Φ_d diagonalisable and Φ_n nilpotent. Moreover, the decomposition of the nilpotent part into the sum of cyclic nilpotent mappings provides the Jordan blocks

$$J_\lambda = \begin{pmatrix} \lambda & 1 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \ddots & \lambda \end{pmatrix} = (J_\lambda)_d + (J_\lambda)_n$$
$$= \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \lambda \end{pmatrix} + \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

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

$$E + t\begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \cdots + \frac{t^{k-1}}{(k-1)!}\begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

where k is the size of the block and E is the identity matrix. The solution of the corresponding matrix system is

$$Y(t) = \mathrm{e}^{\lambda t E}\cdot \mathrm{e}^{Dt} = \mathrm{e}^{\lambda t}\,\mathrm{e}^{Dt} = \begin{pmatrix} \mathrm{e}^{\lambda t} & t\,\mathrm{e}^{\lambda t} & \cdots & \frac{t^{k-1}}{(k-1)!}\mathrm{e}^{\lambda t} \\ 0 & \mathrm{e}^{\lambda t} & \cdots & \frac{t^{k-2}}{(k-2)!}\mathrm{e}^{\lambda t} \\ \vdots & \vdots & \ddots & t\,\mathrm{e}^{\lambda t} \\ 0 & 0 & \cdots & \mathrm{e}^{\lambda t} \end{pmatrix}$$

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.

$$\mathcal{L}(x')(s) = \mathcal{L}(-2x + 3y + 3t^2)(s),$$

$$\mathcal{L}(y')(s) = \mathcal{L}(-4x + 5y + e^t)(s).$$

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

$$s\mathcal{L}(x)(s) - 1 = -2\mathcal{L}(x)(s) + 3\mathcal{L}(y)(s) + \frac{6}{s^3},$$

$$s\mathcal{L}(y)(s) + 1 = -4\mathcal{L}(x)(s) + 5\mathcal{L}(y)(s) + \frac{1}{s-1}.$$

In matrices, this is $\mathbf{A}(s)\hat{\mathbf{x}}(s) = \mathbf{b}(s)$, where

$$\mathbf{A}(s) = \begin{pmatrix} s+2 & -3 \\ 4 & s-5 \end{pmatrix}, \hat{\mathbf{x}}(s) = \begin{pmatrix} \mathcal{L}(x)(s) \\ \mathcal{L}(y)(s) \end{pmatrix} \text{ and } \mathbf{b}(s) = \begin{pmatrix} 1 + \frac{6}{s^3} \\ -1 + \frac{1}{s-1} \end{pmatrix}.$$

Cramer's rule says that

$$\mathcal{L}(x)(s) = \frac{|\mathbf{A}_1|}{|\mathbf{A}|}, \quad \mathcal{L}(y)(s) = \frac{|\mathbf{A}_2|}{|\mathbf{A}|}, \text{ where}$$

$$|\mathbf{A}| = \begin{vmatrix} s+2 & -3 \\ 4 & s-5 \end{vmatrix} = s^2 - 3s + 2,$$

$$|\mathbf{A}_1| = \begin{vmatrix} 1 + \frac{6}{s^3} & -3 \\ -1 + \frac{1}{s-1} & s-5 \end{vmatrix} = (s-5)(1 + \frac{6}{s^3}) + 3(-1 + \frac{1}{s-1}),$$

$$|\mathbf{A}_2| = \begin{vmatrix} s+2 & 1 + \frac{6}{s^3} \\ 4 & -1 + \frac{1}{s-1} \end{vmatrix} = (s+2)(-1 + \frac{1}{s-1}) - 4 - \frac{24}{s^3}.$$

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:

$$x(t) = -\frac{39}{2}t - 3te^t + 28e^t - \frac{21}{4}e^{2t} - \frac{15}{2}t^2 - \frac{87}{4},$$

$$y(t) = -18t - 3te^t + 27e^t - 7e^{2t} - 6t^2 - 21. \ \square$$

N. Numerical solution of differential equations

Now, we present two simple exercises on applying the Euler method for solving differential equations.

Finally, k independent solutions can be written down by inspecting the individual columns in $Y(t)$. Notice that the canonical basis $(e_1, \ldots, e_k)$ provides just the chain of vectors with $D(e_k) = e_{k-1}$, $k = 2, \ldots, k$, while $D(e_1) = 0$. Now the k independent solutions are:

$$y_1(t) = e^{\lambda t} e_1$$

$$y_2(t) = e^{\lambda t}(te_1 + e_2)$$

$$\vdots$$

$$y_k(t) = e^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}e_1 + \frac{t^{k-2}}{(k-2)!}e_2 + \cdots + te_{k-1} + e_k\right)$$

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*

$$y(t) = e^{i\lambda t}\sum_{j=0}^{k} p_j t^j$$

where k is the order of nilpotency of the Jordan block corresponding to the eigenvalue λ 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 578)

$$\Phi'(t) = D^1 X(x_0) \cdot \Phi(t), \quad \Phi(0) = E.$$

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,

$$\Phi(t) = e^{tA}, \quad A = D^1 X(x_0).$$

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 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
$h_0 = 1$
$y^{(0)} = (1\ 0\ 0)$
$h_1 = 0.5$
$y^{(1)} = (1\quad 0.5\quad 0.375\quad 0.3047\quad 0.2583)$
Maximal difference: 0.375.
$h_2 = 0.25$
$y^{(2)} = (1.0000\quad 0.7500\quad 0.6094\quad 0.5165\quad 0.4498$
$\qquad\quad 0.3992\quad 0.3594\quad 0.3271\quad 0.3004)$
Maximal difference: 0.1094.
$h_3 = 0.125$
$y^{(3)} = (1.0000\quad 0.8750\quad 0.7793\quad 0.7034\quad 0.6415$
$\qquad\quad 0.5901\quad 0.5466\quad 0.5092\quad 0.4768\quad 0.4484$
$\qquad\quad 0.4233\quad 0.4009\quad 0.3808\quad 0.3627$
$\qquad\quad 0.3462\quad 0.3312\quad 0.3175)$
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 $y = 1/x$.
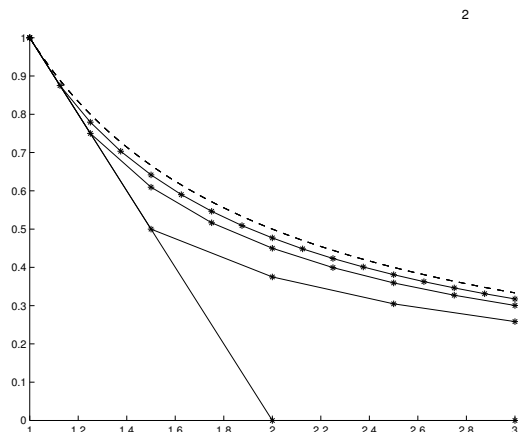
This is a useful step in analysing the qualitative behaviour in a neighbourhood of the stationary point x_0.

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

$$(x_0, y_0) = \left(\frac{\gamma}{\delta\beta}, \frac{\alpha}{\beta}\right)$$

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 determinant of 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 λ (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 β does not appear explicitly here, while δ influences the distortion of the flow lines from being circles. Compare this with the illustrations on page 574. 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,

$$(1\ \ldots\ 1) \cdot A = (1\ \ldots\ 1).$$

Take the exponential e^{tA} to obtain

$$(1\ \ldots\ 1) \cdot e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!}(1\ \ldots\ 1) \cdot A^k = e^t(1\ \ldots\ 1).$$

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.

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 ± 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

$$y_{k+1} = y_k + h \cdot f(x_{k+1}, y_{k+1}).$$

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.

Differentiating with respect to t, yields

$$\frac{d}{dt}B(t) = -\mathrm{e}^{-t}\,\mathrm{e}^{tA}\cdot B_0 + \mathrm{e}^{-t}\,A\,\mathrm{e}^{tA}\cdot B_0 = (-E+A)B(t),$$
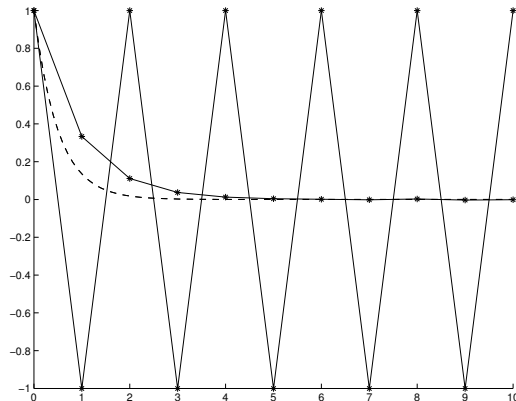
so the matrix $B(t)$ is the solution of the matrix system of equations with constant coefficients

$$Y'(t) = (A - E) \cdot Y(t)$$

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 161. Then its powers converge to a matrix A_∞, 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_∞. 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 ε, there is an $N \in \mathbb{N}$ such that for all $k \geq N$, $\|A^k - A_\infty\| \leq \varepsilon$. Hence,

$$\left\| \mathrm{e}^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k - \mathrm{e}^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A_\infty \right\|$$

$$\leq \mathrm{e}^{-t} \sum_{k<N} \frac{t^k}{k!} C\|A_\infty\| + \mathrm{e}^{-t}\,\varepsilon \|A_\infty\|.$$

Let $t \to \infty$. The limit of the expression $f(t) = \mathrm{e}^{-t} \sum_{k<N} \frac{t^k}{k!}$ 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 ε, there is a time T such that $f(t)$ would be less than ε 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:

Theorem. *Every primitive stochastic matrix A determines a vector system of equations*

$$y'(t) = (A - E) \cdot y(t)$$

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 $t \to \infty$ to the stochastic eigenvector y_∞ 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 6.1.16. 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

$$y(t_{n+1}) = y(t_n) + y'(t_n)h + O(h^2)$$
$$= y(t_n) + f(t_n, y(t_n))h + O(h^2).$$

Define the recurrently the values y_j by the first order formula

$$y_{j+1} = y_j + f(y_j, y_j)h.$$

This leads to the local approximation error $O(h^2)$, occuring 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 then 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}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}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 $2y = x^3 y'$.   ○

8.O.11. Determine all solutions of the equation
$$\sqrt{4 - y^2}\, dx + y\, dy = 0.$$
○

8.O.12. Solve
$$y' \tan x = y^2 + 1 - 2y.$$

8.0.13. Determine the general solution of the differential equation

$$\frac{x^2+1}{x} = \frac{y}{1-y^2}\, y'.$$

8.0.14. Find the general solution of the differential equation

$$(x+1)\, dy + xy\, dx = 0.$$

8.0.15. Find the solution of the differential equation

$$\sin y \cos x\, dy = \cos y \sin x\, dx$$

which satisfies $4y(0) = \pi$.

8.0.16. Solve the initial problem

$$\left(x^2 + 1\right)\left(y^2 - 1\right) + xyy' = 0, \quad y(1) = \sqrt{2}.$$

8.0.17. Determine the particular solution of the equation

$$y' \sin x = y \ln y$$

which goes through the point $[\pi/2, e]$.

8.0.18. Find all solutions of the differential equation

$$2\left(1 + e^x\right) yy' = e^x,$$

which satisfy $y(0) = 0$.

8.0.19. Solve the homogeneous equation

$$(xy' - y)\cos \frac{y}{x} = x.$$

8.0.20. Determine the general solution of the homogeneous differential equation $y^3 = x^3 y'$.

8.0.21. Find all solutions of the equation

$$xy' = \sqrt{x^2 - y^2} + y.$$

8.0.22. Determine the general solution if we are given

$$xy' = y \cos\left(\ln \frac{y}{x}\right).$$

8.0.23. Solve the equation $(x + y)\, dx - (x - y)\, dy = 0$ as homogeneous.

8.0.24. Calculate $y' = (x + y)^2$.

8.0.25. Find the general solution for

$$y' = \frac{x-y+3}{x+y-5}.$$

○

8.0.26. Calculate

$$y' = \frac{x-y+1}{x-y}.$$

○

8.0.27. Determine all solutions of the differential equation

$$y' = \frac{5y-5x-1}{2y-2x-1}.$$

○

8.0.28. 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

$$y' \cos x = (y + 2\cos x)\sin x.$$

597

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

$$3x^2 y' + xy = \frac{1}{y^2}.$$

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

$$xy' + 2y + x^5 y^3 e^x = 0.$$

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

$$y' \left(e^{-y} - x \right) = 1.$$

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

$$y'' + y' + \tfrac{5}{2} \, y = 25 \cos (2x).$$

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

$$y^{(4)} - 2y'' + y = 8\left(e^x + e^{-x}\right) + 4\left(\sin x + \cos x\right).$$

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.O.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

8.C.1. 2.

8.C.2. Factorize the denominator to get $\frac{1}{4}$.

8.C.3. 0.

8.C.4. Expand the fraction with $\frac{y}{y}$ and use the substitution $t = xy$ ($(x, y) \to (0, 2)$, means $t \to 0$). 2.

8.C.5. Use the polar coordinates $x = r \cos \varphi, y = r \sin \varphi$, $(x, y) \to (\infty, \infty)$, means $r \to \infty, \varphi \in (0, \frac{\pi}{2})$.

  0.

8.C.6. Use two different parametrization $y = 1 - e^x$ and $y = x$ to get two different values -4 and 0, thus the limit does not exists.

8.C.7. Try polar cordinates, where $r \to 0, \varphi \in \langle 0, 2\pi)$. The value is $\cos 2\varphi$, that is it depends on the direction of approach to $[0, 0]$, thus the limit does not exists.

8.C.8. Try polar coordinates, where $r \to \infty, \varphi \in (0, \frac{\pi}{2})$. 1 pro $\varphi = \frac{\pi}{4}$, ale pro $\varphi \neq \frac{\pi}{4}$ vyjde 0, takže limita neexistuje.

8.C.9. $\sqrt{2}$.

8.C.10. 2.

8.C.11. 0.

8.C.12. 0.

8.C.13. 0.

8.C.14. 0.

8.C.15. e.

8.C.16. Neexistuje.

8.C.17. Neexistuje.

8.C.18. Choose $y = kx^2$.

8.C.20. The circle $k([0, 0]; 1)$.

8.C.21. The set $\{[x, x + (2k + 1)\frac{\pi}{2}]; x \in \mathbb{R}, k \in \mathbb{Z}\}$.

8.C.22. The function is continuous everywhere including the point $[0, 0]$.

8.D.3. $p = \{[s, \frac{\pi}{4} + \frac{s}{2}, 1 - \pi s]; s \in \mathbb{R}\}$.

8.D.14. a) Compute the differential of the function $f(x, y) = \arctan \frac{x}{y}$ in $[1, 1]$ and get $\frac{\pi}{4} + 0,035$.

  b) $f(x, y) = \ln(x^2 + y^2)$, $P = [x_0, y_0] = [1, 0]$, the result is -0.06.
  c) $f(x, y) = \arcsin \frac{x}{y}$, $x_0 = 0.5, y_0 = 1$, $\frac{\pi}{6} - \frac{0.09}{\sqrt{3}}$.
  d) $f(x, y) = x^y$, $x_0 = 1, y_0 = 2, 1, 08$.

8.D.18. $[4/3, 1/6]$ and $[-1/3, -2/3]$, equations are $x + y = 3/2$ and $x + y = -1$.

8.D.19. $[-7, 1]$ and $[9, -3]$, equations are $x = -7$ and $x = 9$.

8.D.20. $[\sqrt{2}, 1/\sqrt{2}, -1/\sqrt{2}, \sqrt{2}]$ and $[-\sqrt{2}, -1/\sqrt{2}, 1/\sqrt{2}, -\sqrt{2}]$.

8.D.25. $z'_x(1, \sqrt{2}) = \frac{z - 2x}{2z - x - \sqrt{2}y} = 0$, $z'_y(1, \sqrt{2}) = \frac{\sqrt{2}z - 2y}{2z - x - \sqrt{2}y} = 0$, $z''_{xx}(1, \sqrt{2}) = z''_{yy}(1, \sqrt{2}) = -2$, $z''_{xy}(1, \sqrt{2}) = 0$.

8.D.26. $z'_x(-2, 0) = -\frac{4x + 8z}{8x + 2z - 1} = 0$, $z'_y(-2, 0) = -\frac{4y}{8x + 2z - 1} = 0$, $z''_{xx}(-2, 0) = z''_{yy}(-2, 0) = \frac{4}{15}$, $z''_{xy}(-2, 0) = 0$.

8.E.3. As we can easily see, the Taylor polynomial of a function which is a polynomial is the function itself, or the function cut of the higher powers. Therefore, in this case, we have $T(x, y, z) = xz^2 + xy + 1$.

8.E.4. $T(x, y) = y^2$. The tangent plane is given by the linear part of the Taylor polynomial, i. e., $z = 0$. Therefore, the given point does not lie in it.

8.E.5. $T^2_{\ln(xy+1)}(1, 1) = \ln(2) + \frac{1}{4}(x^2 + y^2 + xy - x - y - 1)$.

8.E.6. $y + xy$.

8.F.8. Stationary points: $(\pm\frac{1}{2}, \mp 1, \pm\frac{1}{8})$. The Hessians are indefinite in both cases, no extrema.

8.F.9. Stationary points: $[\mp 2, \pm 1, \pm 2]$. The Hessians are indefinite in both cases, no extrema.

8.F.10. Stationary points: $[\pm\frac{1}{8}, \pm 1, \mp\frac{1}{2}]$. The Hessians are indefinite in both cases, no extrema.

8.F.11. Stationary points: $[\pm 2, \mp 2, \pm 1]$, no extrema.

8.F.12. Stationary points: $(0, -1/4), (\pm\sqrt{3}, -1)$; a minimum occurs at $(0, -1/4)$.

8.F.13. Stationary point: $(0, -1/2)$. The Hessian is indefinite, no extremum.

8.F.14. A global minimum at $(1/7, -2/7)$.

8.F.15. Stationary point: $(-1/9, 2/9)$. The Hessian is indefinite, no extremum.

8.H.17. At the point $[\frac{1}{3}, \frac{1}{6}, -\frac{1}{6}]$, minimum.

8.H.18. At the point $[-\frac{\sqrt{3}}{2}, \frac{3}{2}]$.

8.H.19. $[\frac{3}{2} + \frac{\sqrt{3}}{2}, -\frac{\sqrt[4]{3}}{\sqrt{2}}]$.

8.H.20. $[\frac{3}{2} + \frac{\sqrt{3}}{2}, -\frac{\sqrt[4]{3}}{\sqrt{2}}]$.

8.H.21. At the points $[\pm\frac{1}{\sqrt{2}}, \mp\frac{1}{\sqrt{6}}]$.

8.H.22. At the points $[\pm\frac{\sqrt{6}}{3}, -\frac{\sqrt{3}}{3}]$.

8.H.23. $3\sqrt{3}/16$.

8.H.24. $1/(2\sqrt{6})$.

8.H.25. $(1/sqrt2, 1/2)$, $(-1/sqrt2, -1/2)$.

8.I.23. $[0, \frac{8}{5}]$.

8.I.24. $[\frac{\sqrt{3}}{\pi}, \frac{1}{\pi}]$.

8.I.25. $[0, \frac{4}{3\pi}]$.

8.I.26. $[\frac{\sqrt{3}}{2\pi}, \frac{3}{2\pi}]$.

8.I.27. $V = \pi$.

8.I.28. 8π.

8.I.29. $\frac{\sqrt{2}\pi}{3}$.

8.I.30. $4\sqrt{3}\pi - \frac{16}{3}\pi$.

8.I.32. $\frac{\pi}{6}(17\sqrt{17} - 1)$.

8.I.33. $\sqrt{6}(\pi/4 - 1/2)$.

8.K.7. $y = 1 - \frac{x^2}{4}$, $x \in (0, 2)$.

8.K.8. $Me^{-t/q}$.

8.K.9. $\frac{\sqrt{2}}{2}\sin(2t)$.

8.O.1. $2(100 + 2t) - \frac{15 \cdot 10^5}{(100 + 2t)^2}$.

8.O.2. 800 K.

8.O.3. $-87.5\frac{\ln(0.9)}{\ln(2)} \doteq 13.3$ days.

8.O.4. $\frac{5\ln(5)}{4\ln(2)}$ years.

8.O.5. $\frac{8\ln(3)}{3\ln(2)}$ years.

8.O.6. $5.27\frac{\ln(\frac{4}{3})}{\ln(2)}$.

8.O.7. $y = \frac{x+C}{1-Cx}$. (use the sum formula for the tangent function).

8.O.8. $y \equiv 0$, $y = (1 - Cx)^{-1}$, $C \in \mathbb{R}$.

8.O.9. $y = -\ln(1 - Ce^x)$, $C \in \mathbb{R}$.

8.O.10. $y = Ce^{-1/x^2}$, $C \in \mathbb{R}$.

8.O.11. $y \equiv 2$, $y \equiv -2$, $(x - C)^2 + y^2 = 2^2$, $C \in \mathbb{R}$.

8.O.12. $y \equiv 1$, $y = 1 - \frac{1}{\ln|\sin x| + C}$, $C \in \mathbb{R}$.

8.O.13. $x^2 + 2\ln|x| + \ln|y^2 - 1| = C$, $C \in \mathbb{R}$.

8.O.14. $y = C(x + 1)e^{-x}$, $C \in \mathbb{R}$.

8.O.15. $\sqrt{2}\cos y = \cos x$.

8.0.16. $y = \sqrt{\frac{e^{1-x^2}}{x^2} + 1}$.

8.0.17. $y = e^{\tan{(x/2)}}$.

8.0.18. $y = \pm\sqrt{\ln{(e^x + 1)} - \ln 2}$.

8.0.19. $x = Ce^{\sin{\frac{y}{x}}}$, $C \in \mathbb{R}$.

8.0.20. $y^2 = x^2 + Cx^2y^2$, $C \in \mathbb{R}$.

8.0.21. $y = x$, $y = -x$, $y = x\sin{(\ln{|Cx|})}$, $C \in \mathbb{R} \smallsetminus \{0\}$.

8.0.22. $\cot{\left(\frac{1}{2}\ln{\frac{y}{x}}\right)} = \ln{|Cx|}$, $C \in \mathbb{R} \smallsetminus \{0\}$.

8.0.23. $\arctan{\frac{y}{x}} = \ln{\left(x^2 + y^2\right)} + C$, $C \in \mathbb{R}$.

8.0.24. $y = \tan{(x + C)} - x$, $C \in \mathbb{R}$.

8.0.25. $C = (x - 1)^2 - 2(y - 4)(x - 1) - (y - 4)^2$, $C \in \mathbb{R} \smallsetminus \{0\}$.

8.0.26. $(x - y)^2 + 2x + C = 0$, $C \in \mathbb{R}$.

8.0.27. $y = x$, $C = 5x - 2y + \ln{|y - x|}$, $C \in \mathbb{R}$.

8.0.28. $(x + 1)^2 - 2(x + 1)(y + 2) - (y + 2)^2 = C$, $C \in \mathbb{R} \smallsetminus \{0\}$.

8.0.29. $3(y + 1)^2 - 2(y + 1)(x - 2) + 2(x - 2)^2 = C$, $C \in \mathbb{R} \smallsetminus \{0\}$.

8.0.30. $y = 5 - x + C(x - 3)^2$, $C \in \mathbb{R}$.

8.0.31. $y = Ce^{-3x} + \frac{1}{3}x - \frac{1}{9}$, $C \in \mathbb{R}$.

8.0.32. $y = Ce^{2x} - 3(x + 1)$, $C \in \mathbb{R}$.

8.0.33. $y = \left(x^2 + x + C\right)e^{2x^2}$, $C \in \mathbb{R}$.

8.0.34. $y = \frac{C}{x} + \frac{x\ln x}{2} - \frac{x}{4}$, $C \in \mathbb{R}$.

8.0.35. $y = Cx + x^2\ln x - x^2$, $C \in \mathbb{R}$.

8.0.36. $y = \frac{\sin^2 x + C}{\cos x}$, $C \in \mathbb{R}$.

8.0.37. $y = 3x + \frac{3}{2}e^{-2x} - \frac{3}{2}$, $C \in \mathbb{R}$.

8.0.38. $y = e^{\cos x} + 1$.

8.0.39. $y = \frac{1}{17}\sin x - \frac{4}{17}\cos x + \frac{21}{17}e^{4x}$.

8.0.40. $y = \frac{e^x + ab - e^a}{x}$.

8.0.41. $y^3 = \frac{\ln{|x|} + C}{x}$, $C \in \mathbb{R}$.

8.0.42. $y \equiv 0$, $y^2 = \frac{e^{x^2}}{2x + C}$, $C \in \mathbb{R}$.

8.0.43. $y \equiv 0$, $\frac{1}{y} = \frac{C}{x} + \cos x - \frac{\sin x}{x}$, $C \in \mathbb{R}$.

8.0.44. $y \equiv 0$, $y = x^4\left(\frac{1}{2}\ln{|x|} + C\right)^2$, $C \in \mathbb{R}$.

8.0.45. $y \equiv 0$, $y^{-2} = x^4\left(2e^x + C\right)$, $C \in \mathbb{R}$.

8.0.46. $y^2 + \frac{b}{a} = Ce^{-\frac{2a}{x}}$, $C \in \mathbb{R}$.

8.0.47. $x = \frac{y^2}{2} + Cy^3$, $C \in \mathbb{R}$.

8.0.48. $x = y\ln y + \frac{C}{y}$, $C \in \mathbb{R}$.

8.0.49. $x^2 + y^2\left(y^2 - C\right) = 0$, $C \in \mathbb{R}$.

8.0.50. $x = y\ln y - \frac{y\ln^2 y}{2} + Cy$, $C \in \mathbb{R}$.

8.0.51. $x = (C + y)e^{-y}$, $C \in \mathbb{R}$.

8.0.52. $x = \frac{y^2}{2} + \frac{y}{2} + \frac{1}{4} + Ce^{2y}$, $C \in \mathbb{R}$.

8.0.53. $x = \frac{2}{7}y^3 + \frac{C}{\sqrt{y}}$, $C \in \mathbb{R}$.

8.0.54. $y = C_1e^{-2x} + C_2e^{-x} + (x + 1)e^{3x}$, $C_1, C_2 \in \mathbb{R}$.

8.0.55. E. g., $y = 8\sin{(2x)} - 6\cos{(2x)}$.

8.0.56. $y = \mathrm{e}^{-x}\left(C_1\cos x + C_2\sin x\right) + \frac{3x}{2}\,\mathrm{e}^{-x}\sin x$, $C_1, C_2 \in \mathbb{R}$.

8.0.57.
$$y = \frac{1}{2}e^{(1+\sqrt{2})x} + \frac{1}{2}e^{(1-\sqrt{2}x)x} - 1.$$

8.0.58. $y = \frac{3}{5}e^x - \frac{7}{20}e^{-4x} - \frac{1}{4}$.

8.0.59. $y = C_1\mathrm{e}^x\cos(2x) + C_2\mathrm{e}^x\sin(2x) + \mathrm{e}^{2x}\left(\sin x - \frac{1}{2}\cos x\right)$, where $C_1, C_2 \in \mathbb{R}$.

8.0.60. $y = C_1 + C_2\mathrm{e}^{-x} + \frac{1}{3}x^3 - \frac{3}{2}x^2 + 3x + \mathrm{e}^{2x}$, $C_1, C_2 \in \mathbb{R}$.

8.0.61. $y = \left(C_1 + C_2x\right)\mathrm{e}^x + \left(C_3 + C_4x\right)\mathrm{e}^{-x} + x^2\left(\mathrm{e}^x + \mathrm{e}^{-x}\right) + \cos x + \sin x$, $C_1, C_2, C_3, C_4 \in \mathbb{R}$.

8.0.62. $y = C_1\mathrm{e}^x + C_2x\,\mathrm{e}^x + x\,\mathrm{e}^x\left(\ln|x| - 1\right)$, $C_1, C_2 \in \mathbb{R}$.

8.0.63. $y = C_1\mathrm{e}^{-2x} + C_2x\,\mathrm{e}^{-2x} + \frac{x^2}{2}\,\mathrm{e}^{-2x}\ln x - \frac{3x^2}{4}\,\mathrm{e}^{-2x}$, $C_1, C_2 \in \mathbb{R}$.

8.0.64. For $C_1, C_2 \in \mathbb{R}$, we have $y = -\frac{x}{2}\cos(2x) + \frac{1}{4}\sin(2x)\ln|\sin(2x)| + C_1\cos(2x) + C_2\sin(2x)$.

8.0.65. $y = C_1\cos x + C_2\sin x - 2 + \frac{1}{2}\sin x\ln\left|\frac{1+\sin x}{1-\sin x}\right|$, $C_1, C_2 \in \mathbb{R}$.

8.0.66. $y(x) = -e^{-x} + e^{-\frac{1}{2}x}\sin(\frac{\sqrt{3}}{2}x) + e^{-\frac{1}{2}x}\cos(\frac{\sqrt{3}}{2}x) - \frac{1}{2}\sin(x) - \frac{1}{2}\cos(x)$.

8.0.67. $y = C_1\mathrm{e}^x + C_2\mathrm{e}^{-x} + C_3\mathrm{e}^{2x}$, $C_1, C_2, C_3 \in \mathbb{R}$.

8.0.68. $y = C_1\cos x + C_2\sin x + C_3x\cos x + C_4x\sin x$, where the constants $C_1, C_2, C_3, C_4 \in \mathbb{R}$.

8.0.69. $y = \left(C_1 + C_3x + C_5\mathrm{e}^{-x}\right)\cos x + \left(C_2 + C_4x + C_6\mathrm{e}^{-x}\right)\sin x$, $C_1, C_2, C_3, C_4, C_5, C_6 \in \mathbb{R}$.

8.0.70. $y = C_1 + C_2x + C_3x^2 + C_4\mathrm{e}^{2x} + C_5\mathrm{e}^x + \frac{x^4}{6}$, where the constants $C_1, C_2, C_3, C_4, C_5 \in \mathbb{R}$.

Continuous models – further selected topics

The World might be discrete in reality.

But the continuous models are useful anyhow ...
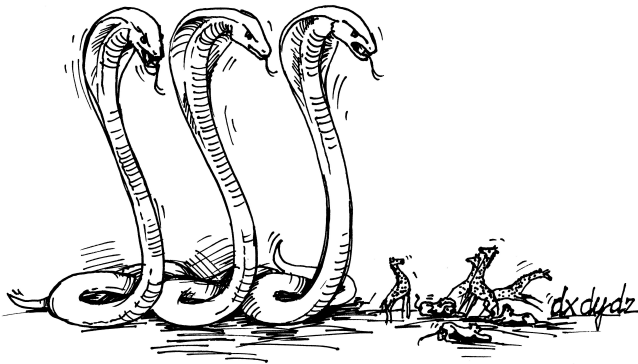
A. Exeterior differential calculus

B. Applications of Stoke's theorem

9.B.1. Compute

$$\int_c (x - y)dx + x \, \mathrm{d}y,$$

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 $\int_c f(x, y) \, \mathrm{d}x + g(x, y) \, \mathrm{d}y$, 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 8.1.23 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 $p : TU \to U$ assigning the foot points to the tangent vectors, we write T_xU 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

$$X(x) = X_1(x)\frac{\partial}{\partial x_1} + \cdots + X_n(x)\frac{\partial}{\partial x_n}.$$

$$= 2 \int_{-2}^{2} \int_{-2}^{2} dx dy = 2 \big[x \big]_{-2}^{2} \cdot \big[y \big]_{-2}^{2} = 32.$$

□

9.B.2. Compute

$$\int_c x^4 \ \mathrm{d}x + xy \ \mathrm{d}y,$$

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:

$$\int_c x^4 dx + xy dy = \iint_D y \ \mathrm{d}x \ \mathrm{d}y = \int_0^1 \int_0^{-x+1} y \ \mathrm{d}x \ \mathrm{d}y$$

$$= \int_0^1 \left[\frac{y^2}{2} \right]_0^{-x+1} \mathrm{d}x = \int_0^1 \left[\frac{x^2 - 2x + 1}{2} \right] \mathrm{d}x$$

$$= \frac{1}{2} \left[\frac{x^3}{3} - \frac{2x^2}{2} + x \right]_0^1 = \frac{1}{6}.$$

□

9.B.3. Calculate

$$\int_c (xy + x + y) \ \mathrm{d}x + (xy + x - y) \ \mathrm{d}y,$$

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

$$\int_c (xy + x + y) \ \mathrm{d}x + (xy + x - y) \ \mathrm{d}y$$

$$= \iint_D y + 1 - x - 1 \ \mathrm{d}x \ \mathrm{d}y$$

$$= \int_0^1 \int_0^{2\pi} r^2 (\sin \varphi - \cos \varphi) \ \mathrm{d}r \ \mathrm{d}\varphi$$

$$= \int_0^1 r^2 \ \mathrm{d}r \int_0^{2\pi} \sin \varphi - \cos \varphi d\varphi$$

$$= \frac{1}{3} \big[- \cos \varphi - \sin \varphi \big]_0^{2\pi} = 0.$$

□

As we know, every differentiable mapping $F : U \to V$ between two open sets $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ defines the mapping $F_* : TU \to TV$ by applying the differential $D^1 F$ to the individual tangent vectors. Thus if $y = F(x) = (f_1(x), \ldots, f_m(x))$ then $F_* : T_x U \to T_{F(x)} V$,

$$F_* \left(\sum_{i=1}^{n} X_i(x) \frac{\partial}{\partial x_i} \right)(y) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \frac{\partial f_j(x)}{\partial x_i} X_i(x) \right) \frac{\partial}{\partial y_j}(y).$$

When we studied the vector spaces in chapter two, we came accros the useful concept of linear forms. They were defined in paragraph 2.3.17 on page 106. 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 $\eta : U \subset \mathbb{R}^n \to T^* U$ with values $\eta(x) \in T_x^* U$ on an open subset U, i.e., $p \circ \eta = \mathrm{id}_U$, we talk about a *differential form* η 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 $(x_1, \ldots, x_n)$ we can use the differentials of the particular coordinate functions to express every linear form η as

$$\eta(x) = \eta_1(x) dx_1 + \cdots + \eta_n(x) dx_n,$$

where $\eta_i(x)$ are uniquely determined functions. Such a form η evaluates on a vector field $X(x) = X_1(x) \frac{\partial}{\partial x_1} + \cdots + X_n(x) \frac{\partial}{\partial x_n}$ as

$$\eta(X)(x) = \eta(x)(X(x)) = \eta_1(x) X_1(x) + \cdots + \eta_n(x) X_n(x).$$

If the form η is the differential of a function f, we get just back the expression

$$X(f)(x) = df(X(x)) = \frac{\partial f}{\partial x_1} X_1(x) + \cdots + \frac{\partial f}{\partial x_n} X_n(x)$$

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 2.3.22 on page 111. Remember also the computation of the volume of parallelepipeds in terms of determinants in 4.1.22 on page 247.

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.

□

9.B.4. Compute $\int\limits_{c} (2e^{2x}\sin y - 3y^3)dx + (e^{2x}\cos y + \frac{4}{3}x^3)dy$, 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

$$x = 3r\cos\varphi, \qquad\qquad \varphi \in [0, 2\pi],$$
$$y = 2r\sin\varphi \qquad\qquad r \in [0, 1],$$

leading to (the Jacobian of the transformation is $6r$):

$$\int\limits_{c} (2e^{2x}\sin y - 3y^3)dx + (e^{2x}\cos y + \frac{4}{3}x^3)\,dy =$$

$$= \iint\limits_{D} 2e^{2x}\cos y + 4x^2 - (2e^{2x}\cos y - 9y^2)\,dx\,dy =$$

$$= \int\limits_{0}^{1}\int\limits_{0}^{2\pi} 6r\left[4(3r\cos\varphi)^2 + 9(2r\sin\varphi)^2\right] =$$

$$= 216\int\limits_{0}^{1} r^3\,dr \int\limits_{0}^{2\pi}\,d\varphi = 216\cdot\left[\frac{r^4}{4}\right]_0^1\cdot 2\pi = 108\pi.$$

$\square$

9.B.5. Compute

$$\int\limits_{c} (e^x\mathrm{ln}y - y^2x)dx + \left(\frac{e^x}{y} - \frac{1}{2}x^2y\right)\,dy,$$

where c is the positively oriented circle $(x-2)^2 + (y-2)^2 = 1$.

Solution.

$$\int\limits_{c} (e^x\mathrm{ln}y - y^2x)dx + (\frac{e^x}{y} - \frac{1}{2}x^2y)dy =$$

$$= \iint\limits_{D} \frac{e^x}{y} - xy - \frac{e^x}{y} + 2xy\,dx\,dy =$$

$$= \int\limits_{0}^{1}\int\limits_{0}^{2\pi} r(r\cos\varphi + 2)\cdot(r\sin\varphi + 2)\,dr\,d\varphi =$$

$$= \int\limits_{0}^{1}\int\limits_{0}^{2\pi} r^3\sin\varphi\cos\varphi + 2r^2(\sin\varphi + \cos\varphi) + 4r\,dr\,d\varphi =$$

$$= \frac{1}{4}\int\limits_{0}^{2\pi} \sin\varphi\cos\varphi\,d\varphi + \frac{2}{3}\int\limits_{0}^{2\pi} \sin\varphi + \cos\varphi\,d\varphi + 4\pi =$$

$$= \frac{1}{3}\left[\frac{\sin^2\varphi}{2}\right]_0^{2\pi} + \left[-\cos\varphi + \sin\varphi\right]_0^{2\pi} + 4\pi = 4\pi.$$

$\square$

Definition. The vector space of all k–linear antisymmetric forms on a tangent space T_xU, $U \subset \mathbb{R}^n$, will be denoted by $\Lambda^k(T_xU)^*$. 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 $G : V \to U$ 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_xV. Just like in the case of linear forms, we can evaluate the form η 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 η* by G.

$$G^*\big(\eta(G(x))\big)(X_1(x),\dots,X_k(x))$$
$$= \eta(G(x))\big(G_*(X_1(x)),\dots,G_*(X_1(x))\big),$$

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^1G. We can compute directly from the definition that, for instance,

$$G^*(dy_i)\big(\frac{\partial}{\partial x_k}\big) = dy_i\big(G_*(\frac{\partial}{\partial x_k})\big) = \frac{\partial g_i}{\partial x_k},$$

and so

$$(1) \qquad G^*(dy_i) = \frac{\partial g_i}{\partial x_1}dx_1 + \cdots + \frac{\partial g_i}{\partial x_m}dx_m,$$

which extends to the linear combinations of all dy_i over fuctions.

Another immediate consequence of the definition is the formula for pullbacks of arbitrary k–forms by composing two diffeomorphisms:

$$(2) \qquad (G \circ F)^*\alpha = F^*\big(G^*\alpha\big).$$

Indeed, as a mapping on k-tuples of vectors,

$$(G \circ F)^*\alpha = \alpha \circ ((D^1G \circ D^1F) \times \dots \times (D^1G \circ D^1F))$$
$$= G^*(\alpha) \circ (D^1F \times \dots \times D^1F) = F^* \circ G^*\alpha$$

as expected.

9.1.3. Wedge product of exterior forms. Given a k–form $\alpha \in \Lambda^k\mathbb{R}^{n*}$ and an ℓ–form $\beta \in \Lambda^k\mathbb{R}^{n*}$, we can create a $(k + \ell)$–form $\alpha \wedge \beta$ by all possible permutations σ of the arguments. We just have to alternate the arguments in all possible orders and take the right sign each time:

$$(\alpha \wedge \beta)(X_1,\dots,X_{k+\ell}) =$$
$$\frac{1}{k!\ell!}\sum_{\sigma\in\Sigma_{k+\ell}}\mathrm{sgn}(\sigma)\alpha(X_{\sigma(1)},\dots,X_{\sigma(k)})\beta(X_{\sigma(k+1)},\dots,X_{\sigma(k+\ell)}).$$

9.B.6. Calculate the integral

$$\int_c (e^x \sin y - xy^2)dx + \left(e^x \cos y - \frac{1}{2}x^2y\right)dy,$$

where c is the positively oriented circle $x^2+y^2+4x+4y+7 = 0$. ◯

9.B.7. Compute

$$\int_c (3y - e^{\sin x}) \, dx + (7x + \sqrt{y^4 + 1}) \, dy,$$

where c is the positively oriented circle $x^2 + y^2 = 9$. ◯

9.B.8. Compute the integral

$$\int_c (\frac{1}{x} + 2xy - \frac{y^3}{3}) \, dx + (\frac{1}{y} + x^2 + \frac{x^3}{3}) \, dy,$$

where c is the positively oriented boundary of the set $D = \{(x,y) \in \mathbb{R}^2 : 4 \le x^2 + y^2 \le 9, \frac{x}{\sqrt{3}} \le y \le \sqrt{3}x\}$. ◯

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

$$m(D) = \frac{1}{2} \int_c -y \, dx + x \, dy =$$

$$= \frac{1}{2} \int_0^{2\pi} a\cos t \cdot b\cos t \, dt - \frac{1}{2} \int_0^{2\pi} b\sin t \cdot (-a\sin t)dt =$$

$$= \frac{1}{2}ab \int_0^{2\pi} \cos^2 t \, dt + \frac{1}{2}ab \int_0^{2\pi} \sin^2 t \, dt =$$

$$= \frac{1}{2}ab \int_0^{2\pi} \cos^2 t + \sin^2 t \, dt = \frac{1}{2}ab2\pi = \pi ab,$$

which is indeed the well-known formula for the area of an ellipse with semi-axes a and b. □

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 α and a k–form β, we get

$$(\alpha \wedge \beta)(X_0, X_1, \ldots, X_k) =$$

$$\sum_{j=0}^{k} (-1)^j \alpha(X_j)\beta(X_0, \ldots, \hat{X}_j, \ldots, X_k),$$

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

$$\varepsilon_{i_1 \ldots i_k} = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

with k-tuples of indices $i_1 < i_2 < \cdots < i_k$ 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 α is given uniquely by functions $\alpha_{i_1 \ldots i_k}(x)$ in the expression

$$\alpha(x) = \sum_{i_1 < \cdots < i_k} \alpha_{i_1 \ldots i_k}(x)dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

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 $G : V \to U$ 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 α is just the multiple of the form α by the function f. Similarly, the top degree forms in $\Omega^n(U)$ are all generated by the single generator $\varepsilon_{12\ldots n}$, since there is just one possibility of n different choices among n coordinates, up to the ordering. This means that actually the n-forms ω are identified with functions via the formula

$$\omega(x) = f(x)dx_1 \wedge \cdots \wedge dx_n.$$

At the same time, while the pullback on the functions $f \in \Omega^0(U)$ by a transformation $F : \mathbb{R}^n \to \mathbb{R}^n$, $y = F(x)$, is trivial, i.e. $F^*f(x) = f(y) = f \circ F(x)$, a straightforward computation reveals

$$(1) \quad F^*\omega(x) = \det(D^1F)(x)f(F(x))dx_1 \wedge \cdots \wedge dx_n$$

for all $\omega = f dy_1 \wedge \cdots \wedge dy_n$.

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

$$m(D) = \tfrac{1}{2} \int_{c_1} -y \ dx + x \ dy + \tfrac{1}{2} \int_{c_2} -y \ dx + x \ dy.$$

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, 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

$$\frac{1}{2} \int_{c_2} -y \ dx + x \ dy = \frac{1}{2} \int_{2\pi}^0 a(t - \sin t) \cdot a(\sin t) \ dt -$$

$$- \frac{1}{2} \int_{2\pi}^0 a(1 - \cos t) \cdot a(1 - \cos t) \ dt =$$

$$= \frac{1}{2} a^2 \int_0^{2\pi} t \sin t - \sin^2 t - 1 + 2 \cos t - \cos^2 t \ dt =$$

$$= \frac{1}{2} a^2 \int_{2\pi}^0 t \sin t + 2 \cos t - 2 \ dt =$$

$$= \frac{1}{2} a^2 [-t \cos t - \sin t + 2 \cos t - 2]_{2\pi}^0 = 3\pi a^2.$$

$\square$

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

$$
\begin{aligned}
x &= \rho \sin \varphi \cos \psi & \rho &= [0, 1], \\
y &= \rho \sin \varphi \sin \psi & \varphi &= [0, \pi], \\
z &= \rho \cos \varphi & \psi &= [0, 2\pi].
\end{aligned}
$$

9.1.4. Integration of exterior forms on $\mathbb{R}^n$. Once we fix coordinates $(x_1, \ldots, x_n)$ on $\mathbb{R}^n$ (e.g. the standard ones), there is the bijection between functions f and top degree forms $\omega(x) = f(x)dx_1 \wedge \cdots \wedge dx_n$. 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 tranformation 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

$$\omega_{\mathbb{R}^n} = dx_1 \wedge \cdots \wedge dx_n.$$

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. $\omega = f(x)dx_1 \wedge \cdots \wedge dx_n$. We define the integral of the form ω as

$$\int_U \omega = \int_U f(x)dx_1 \wedge \cdots \wedge dx_n = \int_U f(x) \ dx_1 \cdots dx_n,$$

where the Riemann integral of a function is considered on the right-hand side.

Let us point out, that the n-form ω on the left-hand side is well defined, independently of any choice of coordinates. If we want to express the form ω in different coordinates using a diffeomorphism $G : V \to U$, $G = (g_1, \ldots, g_n)$, it means we will evaluate ω at a point $G(y) = x$ at the values of the vectors $G_*(X_1), \ldots, G_*(X_n)$. However, this means we will integrate the form $G^*\omega$ in coordinates $(y_1, \ldots, y_n)$, and we already saw in the previous paragraph, cf. 9.1.3(1) that

$$(G^*\omega)(y) = f(G(y)) \det(D^1 G(u))dy_1 \wedge \cdots \wedge dy_n.$$

Substituting into our interpretation of the integral, we get

$$\int_V G^*(f\omega_{\mathbb{R}^n}) = \int_{G^{-1}(U)} f(G(y)) \det(D^1 G(u))dy_1 \cdots dy_n,$$

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 $\mathbb{R}^n$.

The Jacobian of this transformation is $-\rho^2 \sin \varphi$.

The given integral is then equal to

$$I = \iint\limits_S x^3 \, \mathrm{d}y \, \mathrm{d}z \, + y^3 \, \mathrm{d}x \, \mathrm{d}z \, + z^3 \, \mathrm{d}x \, \mathrm{d}y =$$

$$= \iiint\limits_V 3x^2 + 3y^2 + 3z^2 \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \, =$$

$$= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^2 \sin \varphi \left(\rho^2 \sin^2 \varphi \cos^2 \psi + \rho^2 \sin^2 \varphi \sin^2 \psi + \right.$$

$$\left. + \rho^2 \cos^2 \varphi \right) \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\psi =$$

$$= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^4 \sin \varphi \left(\sin^2 \varphi \left(\cos^2 \psi + \sin^2 \psi \right) + \right.$$

$$\left. + \cos^2 \varphi \right) \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\psi =$$

$$= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^4 \sin \varphi \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\psi = 3 \cdot \left[\frac{\rho^5}{5} \right]_0^1 [\psi]_0^{2\pi} [\cos \varphi]_0^\pi =$$

$$= 3 \cdot \frac{1}{5} \cdot 2\pi \cdot [-1 - 1] = -\frac{12}{5} \pi.$$

$\square$

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:

$$\iiint\limits_V \operatorname{div} \vec{F}(x,y,z) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z = \iint\limits_S \vec{F}(x,y,z) \cdot \vec{n}(x,y,z) \, \mathrm{d}S$$

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 $F = (xy^2, yz, x^2 z)$ over the cylinder $x^2 + y^2 = 4, z = 1, z = 3$.

Solution. First of all, we compute the divergence of the vector field:

$$\operatorname{div} F = \nabla \cdot F = \left(\frac{\partial (xy^2)}{\partial x} + \frac{\partial (yz)}{\partial y} + \frac{\partial (x^2 z)}{\partial z} \right) = y^2 + z + x^2.$$

Recall the calculation of the length of a curve in $\mathbb{R}^n$ by univariate integrals, which was discussed in paragraph 6.1.6 on page 378. The curve was parametrized as a mapping $c(t) : \mathbb{R} \to \mathbb{R}^n$, 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 $\rho : \mathbb{R}^n \to \mathbb{R}$ which gave the true size when evaluated at $c'(t)$. This mapping satisfied $\rho(a\,v) = |a|\rho(v)$ since we ignored the orientation of the curve given by our parametrization. If we wanted a signed length, respecting the orientation, then our mapping ρ 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 η 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 $f \circ c$

$$d(f \circ c)(t) = \frac{\partial f}{\partial x_1}(c(t))c_1'(t) + \cdots + \frac{\partial f}{\partial x_n}(c(t))c_n'(t).$$

We can thus try to define the value of the integral in the following way

$$\int_M df = \int_a^b \left(\frac{\partial f}{\partial x_1}(c(t))c_1'(t) + \cdots + \frac{\partial f}{\partial x_n}(c(t))c_n'(t) \right) dt,$$

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

$$\int_{\tilde{a}}^{\tilde{b}} \left(\frac{\partial f}{\partial x_1}(c(\psi(s)))c_1'(\psi(s)) + \cdots \right.$$

$$\left. + \frac{\partial f}{\partial x_n}(c(\psi(s)))c_n'(\psi(s)) \right) \frac{d\psi}{ds} \, ds,$$

and the theorem about coordinate transformations for univariate integrals gives just the same value if we have $\frac{d\psi}{ds} > 0$, i.e.,

Therefore, the flow T of the vector filed is equal to

$$\iiint\limits_{V} y^2 + z + x^2 \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \; =$$

$$= \int\limits_{0}^{2} \int\limits_{0}^{2\pi} \int\limits_{1}^{3} \rho \cdot (\rho^2 \sin^2 \varphi + z + \rho^2 \cos^2 \varphi) \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}z \; =$$

$$= \int\limits_{0}^{2} \int\limits_{0}^{2\pi} \int\limits_{1}^{3} \rho \cdot (\rho^2 (\sin^2 \varphi + \cos^2 \varphi) + z) \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}z \; =$$

$$= \int\limits_{0}^{2} \int\limits_{0}^{2\pi} \int\limits_{1}^{3} \rho \cdot (\rho^2 (\sin^2 \varphi + \cos^2 \varphi) + z) \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}z \; =$$

$$= \int\limits_{0}^{2} \int\limits_{0}^{2\pi} \int\limits_{1}^{3} \rho^3 + \rho z \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}z \; =$$

$$= 2\pi \int\limits_{0}^{2} \int\limits_{1}^{3} \rho^3 + \rho z \, \mathrm{d}\rho \, \mathrm{d}z \; = 2\pi \int\limits_{1}^{3} [\frac{\rho^4}{4} + \frac{\rho^2}{2} z]_0^2 \, \mathrm{d}z \; =$$

$$= 2\pi \int\limits_{1}^{3} 4 + 2z \, \mathrm{d}z \; = 2\pi [4z + z^2]_1^3 = 2\pi [12 + 9 - 4 - 1] = 32\pi$$

$\square$

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:

$$\operatorname{div} F = \nabla \cdot F = (\frac{\partial y}{\partial x} + \frac{\partial x}{\partial y} + \frac{\partial z^2}{\partial z}) = 2z.$$

Thus, the wanted flow equals

$$\iiint\limits_{V} 2z \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z = \int\limits_{0}^{2} \int\limits_{0}^{\pi} \int\limits_{0}^{2\pi} \rho^2 \sin \varphi \cdot 2\rho \cos \varphi \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\psi =$$

$$= 2 \int\limits_{0}^{2} \rho^3 \, \mathrm{d}\rho \int\limits_{0}^{2\pi} \mathrm{d}\psi \int\limits_{0}^{\pi} \sin \varphi \cos \varphi \, \mathrm{d}\varphi =$$

$$= 2 [\frac{\rho^4}{4}]_0^2 \cdot [\psi]_0^{2\pi} \cdot [\frac{\sin^2 \varphi}{2}]_0^{\pi} =$$

$$= 2 \cdot \frac{16}{4} \cdot 2\pi \cdot 0 = 0.$$

$\square$

C. Equation of heat conduction

9.C.1. Find the solution to the so-called equation of heat conduction (equation of diffusion)

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 $\eta = \eta_1 dx_1 + \ldots \eta_n dx_n$ we arrive at the same formulae with η_i replacing the derivatives $\frac{\partial f}{\partial x_1}$,

$$\int_M \eta = \int_a^b \big(\eta_1(c(t))c_1'(t) + \cdots + \eta_n(c(t))c_n'(t)\big) dt,$$

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 η for a general 1–form η. 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 $c(t) : \mathbb{R} \to \mathbb{R}^n$ are the differentiable mappings $\varphi : V \subset \mathbb{R}^k \to \mathbb{R}^n$, $k \leq n$, 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.

<div style="text-align:center">MANIFOLDS IN $\mathbb{R}^n$</div>

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 $\tilde{\varphi} : V \times \tilde{V} \to \mathbb{U}$, $V \subset \mathbb{R}^k$, $\tilde{V} \subset \mathbb{R}^{n-k}$, such that

- the restriction $\varphi = \tilde{\varphi}_{|V} : V \to M$ is an immersion,
- $\tilde{\varphi}^{-1}(M) = V \times \{0\} \subset \mathbb{R}^n$.

The manifolds M are carrying the topology inherited from $\mathbb{R}^n$.

$$u_t(x,t) = a^2\, u_{xx}(x,t), \quad x \in \mathbb{R},\ t > 0$$

satisfying the initial condition $\lim\limits_{t\to 0+} u\,(x,t) = f(x)$.

Notes: The symbol $u_t = \frac{\partial u}{\partial t}$ stands for the partial derivative of the 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 $\frac{\alpha}{c\varrho}$, where α is the coefficient of thermal conductivity, c is the specific heat and ϱ 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

$$\mathcal{F}\,(u_t)\,(\omega,t) = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} u_t(x,t)\, e^{-i\omega x}\, dx =$$

$$= \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} u\,(x,t)\, e^{-i\omega x}\, dx \right)',$$

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

$$\mathcal{F}\,(a^2\, u_{xx})\,(\omega,t) = a^2\, \mathcal{F}\,(u_{xx})\,(\omega,t) =$$
$$-a^2\omega^2\, \mathcal{F}\,(u)\,(\omega,t).$$

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

$$y(\omega,t) = K\,(\omega)\, e^{-a^2\omega^2 t}, \quad K\,(\omega) \in \mathbb{R}.$$

It remains to determine $K(\omega)$. The transformation of the initial condition gives

$$\mathcal{F}\,(f)\,(\omega) = \lim\limits_{t\to 0+} \mathcal{F}\,(u)\,(\omega,t) = \lim\limits_{t\to 0+} y(\omega,t) =$$
$$K\,(\omega)\, e^0 = K\,(\omega),$$

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

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 8.1.23 and the discussion in 8.1.25.

The mapping φ 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 φ 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 $p : TM \to M$.

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.
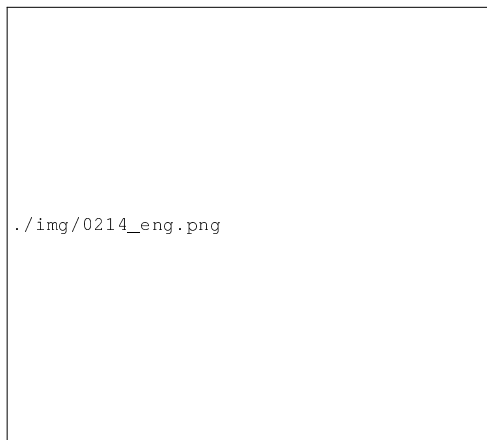
Clearly, every parametrization φ defines a diffeomorphism

$$\varphi_* : TV \to T(\varphi(V)) \subset TM, \quad \varphi_*(c'(t)) = \frac{d}{dt}\varphi(c(t)).$$

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 φ_*.

In particular, the local maps φ (extended to $\tilde{\varphi}$, as in the above definition) induce the local maps $\varphi_* : TU = U \times \mathbb{R}^k \to TM \subset \mathbb{R}^n \times \mathbb{R}^n$ 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.

$$u\left(x,t\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} y(\omega,t)\, e^{i\omega x}\, d\omega =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathcal{F}\left(f\right)(\omega)\, e^{-a^2\omega^2 t}\, e^{i\omega x}\, d\omega =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s)\, e^{-i\omega s}\, ds \right) e^{-a^2\omega^2 t}\, e^{i\omega x}\, d\omega =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s) \left(\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-a^2\omega^2 t}\, e^{-i\omega(s-x)}\, d\omega \right) ds.$$

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)

$$\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-cp^2}\, e^{-irp}\, dp = \frac{1}{\sqrt{2c}} e^{-\frac{r^2}{4c}}, \quad c > 0.$$

According to this formula (consider $c = a^2 t > 0$, $p = \omega$, $r = s - x$), we have

$$\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-a^2\omega^2 t}\, e^{-i\omega(s-x)}\, d\omega = \frac{1}{\sqrt{2a^2 t}} e^{-\frac{(s-x)^2}{4a^2 t}},$$

Therefore,

$$u\left(x,t\right) = \frac{1}{2a\sqrt{\pi t}} \int\limits_{-\infty}^{\infty} f(s)\, e^{-\frac{(x-s)^2}{4a^2 t}}\, ds.$$

$\square$

D. Variational Calculus

E. Complex analytic functions

9.E.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 $f(x+i\,y) = (x + i\,y)^n$ in two real coordinates x, y and see that its derivative is complex.

The first very simple possibility repeats the computation with polynomials in one real variable and is shown in 9.4.1. $\square$

Notice that two differentiable immersions φ and ψ 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 $F : M \to N$ 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 φ of M and ψ 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 $F : M \to N$ defines the tangent mapping $TF : TM \to TN$ 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 $X : M \to TM$ of the footpoint projection $p : TM \to M$, i.e., $p \circ X = \mathrm{id}_M$.

Smooth k–forms η on a manifold M are sections $M \to \Lambda^k(TM)^*$ such that the pullback of this form by any parametrization $V \to M$ 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.[1]

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 ω 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 $\varphi : V \subset \mathbb{R}^k \to U \subset M \subset \mathbb{R}^n$. We consider the standard orientation on $\mathbb{R}^k$ given by the standard basis (cf. 4.1.22 for the definition of the orientation of a vector space). The choice of the parametrization φ 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

---

[1]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 essentialy do not loose any generality here.

is a covering of the entire set M by local parametrizations φ 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 ω 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 ω 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 ψ is a diffeomorphism $\psi : W \to V \subset \mathbb{R}^k$, we can easily compute the result, following the same definition. Let us denote

$$\varphi^*(\omega)(u) = f(u)du_1 \wedge \cdots \wedge du_k.$$

Invoking the relation 9.1.2(2) for the pullback of a form by a composite mapping, we get

$$\int_M \omega = \int_{\mathbb{R}^k} \tilde{\varphi}^*(\omega) = \int_{\mathbb{R}^k} \psi^*(\varphi^* \omega)$$
$$= \int_{\mathbb{R}^k} \psi^*\big(f\, du_1 \wedge \cdots \wedge du_k\big)$$
$$= \int_{\mathbb{R}^k} f(\psi(v)) \det(D^1\psi)(v)dv_1 \cdots dv_k.$$

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

$$M = U_1 \cup U_2 \cup \cdots \cup U_m \cup B,$$

where U_i are pairwise disjoint images of parametrizations φ_i, and B is a set whose inverse image in any parametrization is a Riemann measurable set with measure zero, we can compute

$$\int_M \omega = \int_{U_1} \omega + \cdots + \int_{U_m} \omega,$$

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.

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 φ_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 φ_i.[2]

Now, recall the smooth variants of indicator functions from paragraph 6.1.6. 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 $0 \le f_{\varepsilon,r}(t) \le 1$ everywhere. At the same time, we had $f(t) \ne 0$ 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 $K \ne \mathbb{R}^n$.

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 $B_{r_i}(x_i)$, see the function χ_{r,ε,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

---

[2]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.

$\mathbb{R}^n$. 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. □

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 φ_i. Then, there exists a system of smooth non-negative functions f_i on the sets U_i 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. □

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 $\varphi_i : V_i \to U_i$ 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

$$\int_M \omega = \int_M \sum_i f_i \omega = \sum_i \int_{U_i} f_i \omega,$$

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 φ_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 ω. 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 $\psi : \tilde{V}_j \to \tilde{U}_j$, 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:

$$\sum_i \int_M f_i\omega = \sum_i \int_M f_i\big(\sum_j g_j\big)\omega = \sum_{i,j} \int_M f_i g_j \omega$$

$$\sum_j \int_M g_j\omega = \sum_j \int_M g_j\big(\sum_i f_i\big)\omega = \sum_{i,j} \int_M f_i g_j \omega,$$

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 ω. 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 $d : \Omega^0(\mathbb{R}^n) \to \Omega^1(\mathbb{R}^n)$.

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 $M \subset \mathbb{R}^n$.

<div align="center">Exterior differential</div>

Theorem. *For all m–dimensional manifolds $M \subset \mathbb{R}^n$ and $k = 0, \dots, m$, there is the unique mapping*

$$d : \Omega^k(M) \to \Omega^{k+1}M,$$

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 $\alpha \in \Omega^k(M)$, β 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.

PROOF. Each k–form can be written locally in the form

$$\alpha = \sum_{i_1 < \cdots < i_k} a_{i_1 \cdots i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

If the differential d exists, then by the required properties, it must be equal to

(5)
$$\begin{aligned} d\alpha &= \sum_{i_1 < \cdots < i_k} da_{i_1 \cdots i_k} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \\ &= \sum_{j;\, i_1 < \cdots < i_k} \frac{\partial a_{i_1 \cdots i_k}}{\partial x_j} dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}. \end{aligned}$$

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 $\alpha = a dx_{i_1} \wedge \cdots \wedge dx_{i_k}$ and $\beta = b dx_{j_1} \wedge \cdots \wedge dx_{j_\ell}$, since then the property must hold for the sums of such forms too.

We compute

$$\begin{aligned} d(\alpha \wedge \beta) &= d\big(ab\, dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_\ell} \big) \\ &= \sum_i \frac{\partial a}{\partial x_i} b\, dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_\ell} \\ &\quad + \sum_i \frac{\partial b}{\partial x_i} a\, dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_\ell} \\ &= d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta, \end{aligned}$$

as expected.

The last property can be again verified for simple forms $\alpha = a\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}$. Applying the formula (5) twice, we arrive at

$$\begin{aligned} d(d\alpha) &= d\left(\sum_i \frac{\partial a}{\partial x_i} dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \right) \\ &= \sum_j \sum_i \frac{\partial^2 a}{\partial x_i \partial x_j} dx_j \wedge dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \\ &= \sum_{i<j} \Big(\frac{\partial^2 a}{\partial x_i \partial x_j} - \frac{\partial^2 a}{\partial x_j \partial x_i} \Big) dx_j \wedge dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \\ &= 0, \end{aligned}$$

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.

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.[3]

Thus, consider a change of coordinates $G : U \to V$, $x = G(y) = (g_1(y), \ldots, g_m(y))$, and compute $G^*(d\alpha)$ of an exterior form $\alpha = a\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}$ (which gives the result for sums of such expressions too). This is straightforward:

$$G^*(d\alpha) = \sum_i G^*\left(\frac{\partial a}{\partial x_i}\right) G^*(dx_i) \wedge G^*(dx_{i_1}) \wedge \cdots$$

$$= \sum_i \left(\frac{\partial a}{\partial x_i} \circ G\right)\left(\frac{\partial g_i}{\partial y_1} dy_1 + \cdots\right) \wedge \left(\frac{\partial g_{i_1}}{\partial y_1} dy_1 + \cdots\right) \wedge \cdots.$$

Now, notice $d(G^*(dx_j)) = G^*(d(dx_j)) = 0$ and thus

$$d\big(G^*(\alpha)\big) = d\big((a \circ G) G^*(dx_{i_1}) \wedge \ldots\big)$$

$$= d(a \circ G) \wedge G^*(dx_{i_1}) \wedge \cdots \wedge G^*(dx_{i_k})$$

$$= \sum_i \left(\left(\frac{\partial a}{\partial x_i} \circ G\right)\left(\frac{\partial g_i}{\partial y_1} dy_1 + \ldots\right)\right) \wedge G^* dx_{i_1} \wedge \ldots,$$

clearly the same expressions. □

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 ∂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 $c : [a, b] \to M$, then we get directly from the definition that

$$\int_M df = \int_{[a,b]} c^*(df) = \int_a^b \frac{d}{dt}(f \circ c)(t)\, dt$$

$$= f(c(b)) - f(c(a)).$$

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

---

[3]Such operators are intrinsically defined on all manifolds. Actually, for all $k > 0$, the only operation $d : \Omega^k \to \Omega^{k+1}$ commuting with pullbacks and with values depending only on the behavior of the argument α 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. Slovak for full proof of this astonishing claim.

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 $\partial M = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n; \ x_1 = 0\}$. The orientation on this boundary inherited from the standard orientation is the one determined by the form $dx_2 \wedge \cdots \wedge dx_n$.

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 φ_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 $\varphi : V \subset (-\infty, 0] \times \mathbb{R}^{m-1} \to M$ 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 ∂M defines the structure of an $(m-1)$–dimensional oriented manifold on ∂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.

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 ∂M. Thus we shall not go into details of this technical extension of our definitions.

pridat obrazek k prikladu

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 ω with compact support on an oriented manifold $\bar{M}$ with boundary ∂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 ω is compact) of integrals of forms ω 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, ω will surely be the sum of forms ω_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 ω_j separately. Let us compute the integrals $\int_M d\omega_j$ using the Fubini's theorem. This is most simple if $\bar{M} = \mathbb{R}^n$,

$$\int_{\mathbb{R}^n} d\omega_j = (-1)^{j-1}\int_{\mathbb{R}^{k-1}}\Big(\int_{-\infty}^{\infty}\frac{\partial a_j}{\partial x_j}\,dx_j\Big)dx_1 \cdots \hat{dx_j} \cdots dx_k$$

$$= (-1)^{j-1}\int_{\mathbb{R}^{k-1}}\big[a_j\big]_{-\infty}^{\infty}dx_1 \cdots \hat{dx_j} \cdots dx_k = 0.$$

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 ω_j are all zero outside of such a large interval I and thus

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 ω_j evaluates identically to zero on the
boundary ∂M, since x_1 is constant there and thus dx_1 is iden-
tically zero on all tangent directions to ∂M. Integration over
the interior M yields zero, using the same approach as above:

$$\int_M d\omega_j = (-1)^{j-1} \int_{-\infty}^0 \int_{\mathbb{R}^{k-2}} \left(\int_{-\infty}^{\infty} \frac{\partial a_j}{\partial x_j}\, dx_j \right) dx_1 \cdots \hat{dx_j} \cdots dx_k$$

$$= (-1)^{j-1} \int_{-\infty}^0 \int_{\mathbb{R}^{k-1}} \big[a_j \big]_{-\infty}^{\infty} dx_1 \cdots \hat{dx_j} \cdots dx_k = 0$$

since the function a_j has compact support. So the theorem is
also true in this case.

However, if $j = 1$, then we obtain

$$\int_M d\omega_1 = \int_{\mathbb{R}^{k-1}} \left(\int_{-\infty}^0 \frac{\partial a_1}{\partial x_1}\, dx_1 \right) dx_2 \cdots \cdots dx_k$$

$$= \int_{\mathbb{R}^{k-1}} a_1(0, x_2, \ldots, x_k) dx_2 \cdots dx_k = \int_{\partial M} \omega_1.$$

This finishes the proof of Stokes' theorem. □

9.1.13. Green's theorem. We have proved an extraordinar-
ily strong result which covers several standard
integral relations from the classical vector anal-
ysis. 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 ∂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 ω_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 ω_X over a curve C has the physical inter-
pretation 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

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.

<div align="center">GAUSS–OSTROGRADSKY'S THEOREM</div>

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

$$\int_S f\, dy \wedge dz + g\, dz \wedge dx + h\, dx \wedge dy$$

$$= \int_M \left(\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z} \right) dx \wedge dy \wedge dz.$$

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 particuclar, if $\operatorname{div} X$ vanished identically, then the total volume flow through the boundary surface of the region is zero as well.

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 ω_S on S is defined by the property $\nu^* \wedge \omega_S = \omega_{\mathbb{R}^3}$ at all points of S, where ν^* is dual form to the oriented (outward) unit normal to S.

All forms of degree 2 are multiples of ω_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 folowing 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.

<div align="center">DIVERGENCE THEOREM</div>

Theorem. *Let X be a vector field on a n-dimensional manifold $M \subset \mathbb{R}^n$ with hypersurface boundary S. Then*

$$\int_M \operatorname{div} X \, dx_1 \ldots dx_n = \int_S X \cdot \nu \, dS,$$

where ν 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 ω 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*:

<div align="center">THE CLASSICAL STOKES' THEOREM</div>

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 $\omega_X = f \, dx + g \, dx + h \, dz$. 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 $\operatorname{rot} X$,

$$d\omega_X = \omega_{\mathbb{R}^3}(\operatorname{rot} X, \, , \,).$$

This field is called the *rotation* or *curl* of the vector field X. The Stokes' theorem now reads:

$$\int_C \omega_X = \int_M \operatorname{rot} X.$$

Consequently, the fields X with the property $\omega_X = dF$ for some function F (the fields of gradients of functions), have got the property $\operatorname{rot} X = 0$. They are called *conservative* (or *potential*) vector fields.

9.1.16. 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 paramaterized 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 $X_1, \ldots, X_n$ 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 neccesary 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 $X_1, \ldots, X_n$ in order to "draw" new coordinates, in which the integral submanifold would appear as given by $x_{n+1} = 0, \ldots, x_m = 0$. The problem we face is that the flows do not commute in general and thus our idea will not work.

9.1.17. Lie bracket of vector fields. Fortunately, the commutativity of the flows is captured by a simple differential operation.

Consider two vector fields on $\mathbb{R}^m$, $X = X_1(x)\frac{\partial}{\partial x_1} + \cdots + X_m(x)\frac{\partial}{\partial x_m}$, $Y = Y_1(x)\frac{\partial}{\partial x_1} + \cdots + Y_m(x)\frac{\partial}{\partial x_m}$. The commutator of the derivatives of functions in the directions

of these vector fields is

$$Y(Xf) - X(Yf) = \sum_{i,j} Y_i \frac{\partial}{\partial x_i}\left(X_j \frac{\partial f}{\partial x_j}\right) - X_i \frac{\partial}{\partial x_i}\left(Y_j \frac{\partial f}{\partial x_j}\right)$$

$$= \sum_{i,j}\left(Y_j \frac{\partial X_i}{\partial x_j} - X_j \frac{\partial Y_i}{\partial x_j}\right)\frac{\partial f}{\partial x_i},$$

thanks to the commutativity of the second derivatives of f. Thus, the commutator of the two vector fields behaves as the vector field $[X, Y]$,

$$[X, Y] = \sum_{i,j=1}^{m}\left(Y_j \frac{\partial X_i}{\partial x_j} - X_j \frac{\partial Y_i}{\partial x_j}\right)\frac{\partial}{\partial x_i}.$$

This vector field is called the *Lie bracket*[4] 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 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.

9.1.18. 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 $i : N \to M$ 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 $i(N) \subset$

[4]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 algbebras 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.

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:

Claim. If $\varphi : N \to M$ 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 φ 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 $X_1 \ldots, X_n$. Actually, we shall prove a much stronger version of the theorem. We claim that if D is involutive, then there are coordinates $(x_1, \ldots, x_m)$ around each point $x \in M$, such that the equations $x_{n+1} = a_{n+1}, \ldots, x_m = a_m$ 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 $y_1, \ldots, y_m$ 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 $y_2 = 0, \ldots, y_m = 0$ and the "projections" Y_j of the other fields to make then tangent to Q. This requires that Y_j leave constant the coordinate y_1, i.e. we set

$$Y_j = X_j - X_j(y_1)X_1, \quad j = 2, \ldots, m.$$

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 $Y_1, \ldots, Y_n$ 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 $y_2 = b_2, \ldots, y_m = b_m$ with small constants b_i and there is the projection $p : U \to Q$ forgetting the first coordinate.

On the submanifold Q, there is the $(n-1)$-dimensional involutive distribution $\tilde{D}$ generated by the fields $\tilde{Y}_i = Y_{i|Q}$, $i = 2, \ldots, n$ (notice we again use the argument from the beginning of the proof about the brackets of restricted fields). Now, our assumption says we find suitable coordinates $(q_2, \ldots, q_m)$ on Q around the point $x \in Q$, so that for all small constants $b_{n+1}, \ldots, b_m$, the integral submanifolds of $\tilde{D}$ are defined by $q_{n+1} = b_{n+1}, \ldots, q_m = b_m$.

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,

$$ x_1(y) = y_1(y), x_2(y) = q_2(p(y)), \ldots, x_m = q_m(p(y)). $$

The hope is that all submanifolds N given by equations $x_{n+1} = b_{n+1}, \ldots, x_m = b_m$ (for small b_j) will be tangent to all fields $Y_1, \ldots, Y_n$. Technically, this means $Y_i(x_j) = 0$ for all $i = 1, \ldots, n$, $j = n+1, \ldots, m$. 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

$$ \frac{\partial}{\partial x_1}(Y_i(x_j)) = Y_1(Y_i(x_j)) = Y_i(Y_1(x_j)) + [Y_1, Y_i](x_j) $$

$$ = Y_i(Y_1(x_j)) + c_{1i1}Y_1(x_j) + \sum_{k=2}^{m} c_{1ik}Y_k(x_j) $$

$$ = \sum_{k=2}^{m} c_{1ik}Y_k(x_j). $$

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. $\qquad\square$

9.1.19. 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 ω_j on M.

A direct computation in coordinates reveals that the differential of linear from ω 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

$$X(\omega(Y)) - Y(\omega(X)) = \sum_{i,j} (X_i \frac{\partial}{\partial x_i}(\omega_j Y_j) - Y_i \frac{\partial}{\partial x_i}(\omega_j X_j))$$

$$= \sum_{i,j} \left(X_i \frac{\partial \omega_j}{\partial x_i} Y_j - Y_i \frac{\partial \omega_j}{\partial x_i} X_j + \omega_j \left(X_i \frac{\partial Y_j}{\partial x_i} - Y_i \frac{\partial X_j}{\partial x_i} \right) \right)$$

$$= d\omega(X,Y) + \omega([X,Y]).$$

Thus, the involutivity of a distribution defined by linear forms $\omega_{n+1}, \ldots, \omega_m$ should be closely linked to properties of the differentials on the common kernel. Indeed, there is the following version of the latter theorem:

<center>FROBENIUS' THEOREM</center>

Theorem. *The distribution D defined on an m-dimensional manifold M by $(m-n)$ independent smooth linear forms $\omega_{n+1}, \ldots, \omega_m$ is integrable if and only if there are linear forms α_{ij} such that $d\omega_k = \sum_\ell \alpha_{k\ell} \wedge \omega_\ell$.*

PROOF. Let us write $\omega = (\omega_{n+1}, \ldots, \omega_m)$ for the $\mathbb{R}^{m-n}$-valued form. The distribution is $D = \ker \omega$. Now, the formula (1) (applied to all components of ω) 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 ω 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 $(x_1, \ldots, x_m)$ such that D is the common kernel of all $dx_{n+1}, \ldots, dx_m$. In particular, our forms ω_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

$$dx_k = \sum_\ell a_{k\ell} \omega_\ell, \quad k, \ell = n+1, \ldots, m.$$

Finally, $d\omega_k$ includes only terms with $dx_i \wedge dx_j$ with $j > n$ and all dx_j can be expresed via our forms ω_ℓ from the previous equation. Thus the differentials have got the requested forms. $\square$

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.

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 explicitely, we dealt with vector equations

$$x^{(k)} = F(t, x, \dot{x}, \ddot{x}, \dddot{x}, \dots, x^{(k-1)}), \quad F : \mathbb{R}^{nk+1} \to \mathbb{R}^n,$$

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 derivates $x_j = x^{(j)}$ 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 sence 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

$$(1) \qquad a(x, y)u_x + b(x, y)u_y = 0$$

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 $\Omega \times \mathbb{R} \subset \mathbb{R}^3$. Usualy, Ω is an open set together with a nice boundary, a curve $\partial\Omega$ in our case.

An obvious simple idea suggests to write Ω 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 Ω. 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) $$\dot{x} = a(x, y), \qquad \dot{y} = b(x, y).$$

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 of its *characteristic equations* (2). If the coeficients 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

(3) $$yu_x - xu_y = 0, \qquad u(x, 0) = x.$$

We saw already the solutions to the characteristic equations

$$\dot{x} = y, \qquad \dot{y} = -x$$

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)

(1) $$a(x, y, u)u_x + b(x, y, u)u_y = f(x, y, u).$$

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) $$\dot{x} = a(x, y, z), \quad \dot{y} = b(x, y, z), \quad \dot{z} = f(x, y, z).$$

This autonomous system of three equations is uniquely solvable for each initial condition if a, b, and f are Lipschitz.

A function ψ 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)) = $ 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 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) $$u(x, y) = h(x^2 + y^2)$$

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 intital values to an open neighborhood of $\partial\Omega$.

Of course, we may similarly use characteristics and integrals for any finite number of variables $x = (x_1, \ldots, x_n)$ and equations of the form

$$a_1(x, u)\frac{\partial u}{\partial x_1} + \cdots + a_n(x, u)\frac{\partial u}{\partial x_n} = f(x, u)$$

with the unknown function $u = u(x_1, \ldots, x_n)$. 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 $u = (u_1, \ldots, u_m) : \mathbb{R}^n \to \mathbb{R}^m$, subject to m equations

(1) $$A_i(x, u) \cdot \nabla u_i = f_i(x, u), \quad i = 1, \ldots m,$$

where the left hand side means the scalar product of a vector valued function $A_i : \mathbb{R}^{m+n} \to \mathbb{R}^n$ 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 ∇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 $F(x, u) : \Omega \times \mathbb{R} \to \mathbb{R}^m$, we can consider the system of equations

(2) $$A(x,u) \cdot \nabla u = F(x,u).$$

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 neccesary conditions from the individual characteristic fields.

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) $$\nabla u = A^{-1}(x,u) \cdot F(x,u) = G(x,u).$$

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 ∇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) $$\frac{\partial u_p}{\partial x_i} = F_{pi}(x,u), \quad i = 1, \ldots, n, \ p = 1, \ldots, m.$$

The necessary conditions imposed by the symmetry of higher order derivatives then read

(2) $$\frac{\partial^2 u_p}{\partial x_i \partial x_j} = \frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} = \frac{\partial F_{pj}}{\partial x_i} + \sum_q \frac{\partial F_{pj}}{\partial u_q} F_{qi}$$

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 $u : \mathbb{R}^n \to \mathbb{R}^m$. The equations (1) describe the n-dimensional distribution D on $\mathbb{R}^{m+n}$ and the graphs of possible solutions $u = (u_1, \ldots, u_m)$ are just the integral manifolds of D. The distribution D is clearly defined by the m linear forms

$$\omega_p = du_p - \sum_i F_{pi} dx_i, \quad p = 1, \ldots, m,$$

while the vector fields generating the common kernel of all ω_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

$$-d\omega_p = \sum_{i,j} \frac{\partial F_{pi}}{\partial x_j} dx_j \wedge dx_i + \sum_{i,q} \frac{\partial F_{pi}}{\partial u_q} du_q \wedge dx_i$$

$$= \sum_{i,j} \Big(\frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} \Big) dx_j \wedge dx_i$$

$$-d\omega_p(X_j, X_i) = \Big(\frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} \Big)$$

$$- \Big(\frac{\partial F_{pj}}{\partial x_i} + \sum_q \frac{\partial F_{pj}}{\partial u_q} F_{qi} \Big).$$

Thus, vanishing of the differentials on the common kernel is equivalent to the neccesary 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 obstractions 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) $Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = 0$

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 α, β, we arrive at

$$(2) \quad (A\alpha^2 + 2B\alpha\beta + C\beta^2)f'' + (D\alpha + E\beta)f' + Ff = 0.$$

This is a nice ODE as soon as we fix the values of α and β. Let us look at several simple cases of special importance.

Assume $D = E = F = 0$, $A \neq 0$. Then, after dividing by α^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 $x - cy =$ 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 $f : \mathbb{C} \to \mathbb{C}$ viewed as a mapping on the complex plane. Remind that some of such mappgins 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 $\mathbb{C}$ holomorhic 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 begining 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 $f(x + i\,y) + g(x - i\,y)$ will be real general solutions.

For example, consider $f(p) = p^2$ leading to

$$u(x, y) = (x + iy)^2 = (x^2 - y^2) + i\,2xy$$

and simple check shows that both terms satisfy the equation separately. Notice the two solutions $x^2 - y^2$ and xy povide 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 ν 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 $\nu = \frac{\beta}{\kappa \alpha^2}$:

$$u(x, y) = f(p) = e^{\frac{1}{\kappa}\left(\frac{\beta}{\alpha}x + \frac{\beta^2}{\alpha^2}y\right)}.$$

Again, a simple check reveals that this is a solution. But it is not very "general" – it depends just on two scalars α and β. 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 differenatial 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

$$u(x, y) = \frac{1}{6}(x^3 + y^3) + f(x - i\,y) + g(x + i\,y)$$

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. λ^2, there will be a sign change in our equations and the resulting general solution is

$$u(x,t) = (A\cosh\alpha x + B\sinh\alpha x)\,\mathrm{e}^{\alpha^2\kappa t}\,.$$

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 α^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

$$T(t) = A\cos c\alpha t + B\sin c\alpha t$$
$$X(x) = C\cos\beta x + D\sin\beta x$$
$$Y(y) = E\cos\gamma y + F\sin\gamma y$$
$$Z(z) = G\cos\delta z + H\sin\delta z$$

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. $X(x) = \mathrm{e}^{\pm i\beta x}$ or $X(x) = \mathrm{e}^{\pm\beta x}$, etc). For instance, take one of the solutions with all the separation constants negative

$$u(x,y,z,t) = \mathrm{e}^{i\beta x}\,\mathrm{e}^{i\gamma y}\,\mathrm{e}^{i\delta z}\,\mathrm{e}^{-ic\alpha t} = \mathrm{e}^{i(\beta x + \gamma y + \delta z - c\alpha t)}\,.$$

Similarly to the 1D situation, we can again see a "plane wave" propagating along the direction (β, γ, δ) with angular frequency $c\alpha$.

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 Ω. We write $\partial\Omega$ for the boundary of Ω and assume this is a smooth manifold.

Locally, such a submanifold in $\mathbb{R}^n$ is given by one implicit function $\rho : \mathbb{R}^n \to \mathbb{R}$ 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 Ω, if it is differentiable on its interior and the directional derivatives $D^1_\nu u(x)$ 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 $Lu = F(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$.

<div align="center">CAUCHY BOUNDARY PROBLEM</div>

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 Ω, 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.

<div align="center">DIRICHLET AND NEUMANN BOUNDARY PROBLEMS</div>

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 Ω, subject to the condition $u = \varphi$ on $\partial\Omega$.

The *Neumann problem* is to solve the equation $Lu = F$ on Ω, 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

$$(1) \qquad \Delta u = \left(\frac{\partial^2}{\partial x_1{}^2} + \cdots + \frac{\partial^2}{\partial x_n{}^2} \right) u = F(x_1, \ldots, x_n).$$

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 $\partial\Omega$

$$(2) \qquad \int_M \operatorname{div} X \, dx_1 \ldots dx_n = \int_{\partial\Omega} X \cdot \nu \, d\partial\Omega,$$

where ν 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 φ and ψ. Then*

$$(3) \qquad \int_M \left(\varphi \Delta\psi + \nabla\varphi \cdot \nabla\psi \right) dx_1 \ldots dx_n = \int_S \varphi \nabla\psi \cdot \nu \, dS.$$

This version of the divergence theorem is called the 1st Green's identity.

Next, let us consider one more differentiable function μ and $X = \varphi\mu\nabla\psi - \psi\mu\nabla\varphi$. The the divergence theorem yields the so called 2nd Green's identity

$$(4) \qquad \begin{aligned} \int_M & \varphi(\nabla \cdot (\mu\nabla))\psi - \psi(\nabla \cdot (\mu\nabla))\varphi \, dx_1 \ldots dx_n = \\ & \int_S \mu(\varphi\nabla\psi - \psi\nabla\varphi) \cdot \nu \, dS, \end{aligned}$$

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 φ and ψ are differentiable functions and $\nabla\psi$ is the gradient of ψ. Indeed,

$$i_X \omega_{\mathbb{R}^n} = \varphi(\nabla\psi \cdot \nu) dS$$
$$\operatorname{div} X = \varphi \Delta\psi + \nabla\varphi \cdot \nabla\psi,$$

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 ψ in the direction of the oriented unit normal ν.

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 ψ and φ vanish on the boundary $\partial\Omega$, we obtain

$$\int_\Omega \varphi\Delta\psi - \psi\Delta\varphi \, dx_1 \ldots dx_n = 0.$$

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 Ω, 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 $\varphi = \psi = u$,

$$\int_\Omega (u\Delta u + \nabla u \cdot \nabla u) \, dx_1 \ldots dx_n = \int_{\partial\Omega} u\frac{\partial u}{\partial n} dS.$$

In both problems, Dirichlet or Neumann, the right hand side vanishes. The first term in the left hand integrand vanishes, too. We conclude

$$\int_\Omega \|\nabla u\|^2 \, dx_1 \ldots dx_n = 0,$$

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 coinside on the boundary and thus they are equal. □

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_α and B_α depending on the chosen frequency α. 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 α towards infinity, the Cauchy boundary conditions can become arbitrarily small and still small change of B_α 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.

The problem $Lu = F$ on the domain Ω 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.

Usualy, 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 uniquenes 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 plynomials 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, ∇u means again the gradient vector of u.

For unknown scalar function u on a domain $\Omega \subset \mathbb{R}^n$ we prescribe its derivatives

$$(1) \quad \sum_{|\alpha|=k} a_\alpha(x, u, \ldots, \nabla^{k-1}u)\partial_\alpha u = b(x, u, \ldots, \nabla^{k-1}u),$$

where b and a_α are functions on the tubular domain $\Omega \times \mathbb{R}^N$, accomodating all the derivatives, with at least one of a_α non-zero. We call such equations the (scalar) *quasilinear partial differential equations* (PDE) of order r.

We call (1) semilinear if all a_α 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 Ω,

$$P(x) : (\mathbb{R}^n)^k \to \mathbb{R}, \quad P(x, \xi, \ldots, \xi) = \sum_{|\alpha|=k} a_\alpha(x)\xi^\alpha.$$

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 $\Box u = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0$ 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 $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$, 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 kth-order derivatives of the components of u, $b : \Omega \times \mathbb{R}^N \to \mathbb{R}^m$, 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. Γ 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 Γ determining a solution of 9.2.12(1), at least locally around a given point.

To make things easy, let us first assume that Γ 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 Γ, we also now all derivatives ∂_α with $\alpha = (\alpha_1, \dots, \alpha_{n-1}, 0)$, $0 \le |\alpha|$. Thus, we have to choose reasonably differentiable functions c_j on Γ, $j = 0, \dots, k-1$, 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 Γ, $0 \le |\alpha| < \infty$ with $\alpha_n < k$ are computed inductively by the symmetry of partial derivatives.

Moreover, if $a_{(0,\dots,0,k)} \ne 0$, 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 $a = a_{(0,\dots,0,k)}(x, u, \dots, \nabla^{k-1}(u)) \ne 0$ (and similarly leaving out the arguments of the other functions a_α), the equation 9.2.12(1) can be rewritten as

(1) $\quad \dfrac{\partial^k}{\partial x_n{}^k} u = \dfrac{1}{a}\Big(-\sum_{|\alpha|=k, \alpha_n \ne k} a_\alpha \partial_\alpha u + b(x, u, \dots, \nabla^{k-1}u)\Big)$.

Now, on Γ 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 $D^1_{\nu(x)}u$ of u in the direction of the unit normal vector ν to the hypersurface Γ:

Cauchy data for scalar PDE

The (smooth or analytic) *Cauchy data* for the kth order quasi-linear PDE 9.2.12(1) consist of a hypersurface $\Gamma \subset U$ and k (smooth or analytic) functions $c_j, 0 \le j \le k-1$, prescribing the derivatives in the normal directions to Γ

$$(2) \qquad (D^1_{\nu(x)})^j u(x) = c_j(x), \quad x \in \Gamma.$$

A normal direction $\nu(x), x \in \Gamma$, is called *characteristic* for the given Cauchy data, if

$$(3) \qquad \sum_{|\alpha|=k} a_\alpha(x, u, \dots, \nabla^{k-1} u)\nu(x)^\alpha = 0.$$

The Cauchy data are called *non-characteristic* if there are no characteristic normals to Γ.

Notice the situation simplifies for the semi-linear equations. Then the characteristic directions do not depend on the chosen functions c_j from the Caychy 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 $a_{(0,\dots,0,k)} \ne 0$.

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 Γ. 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 Γ are non-characteristic if and only if $(T\Gamma)^\perp \cap \mathcal{C} = \{0\}$, i.e. the orthogonal complements to the tangent spaces to Γ 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 6.2.15 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 Γ 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 kth 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.

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 Γ.*

Notice that we have computed explicitly the formal power series for the solution (by an inductive procedure) for the special case when Γ 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 Γ discussed in the beginning of 9.2.13. Remind that for such Γ the non-characteristic condition in 9.2.13(3) reads $a_{(0,...,0,k)} \neq 0$.

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 $\ell = 0, \ldots, k - 1$)

(1) $$\sum_{|\alpha|=k} a_\alpha \partial_\alpha u = b, \quad \frac{\partial^\ell u}{\partial n^\ell}(x) = c_\ell(x), \ x \in \Gamma.$$

We shall work locally around some unspecified fixed point in Γ. Since Γ 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, Ψ can be chosen again analytic. Thus, the unit normal vector ν to Γ equals to $\nabla \Psi_n$, up to a multiple μ^{-1} at each point of Γ.

Let $\Phi = \Psi^{-1}$, i.e. $x = \Phi(y)$, and $v(y) = u(\Phi(y))$. Then Φ is analytic and the equation transforms to another equation in the coordinates y,

(2) $$\sum_{|\alpha|=k} \tilde{a}_\alpha \partial_\alpha v = \tilde{b}$$

with analytic coefficients (they can be all expressed by means of the chain rule and the mutually inverse transformations Φ, Ψ). By the very definition, $\frac{\partial \Phi}{\partial y_n}$ is a vector (the last column in the matrix $D^1 \Phi$) perpendicular to Γ 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 $\nabla \Psi_j$, $j = 1, \ldots, n-1$ 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 $\tilde{\Gamma}$.

$$v = c_o \circ \Phi$$

$$\frac{\partial v}{\partial y_n} = \nabla u \cdot \frac{\partial \Phi}{\partial y_n} = \mu \nabla u \cdot \nu = \mu \frac{\partial u}{\partial \nu} = \mu c_1$$

$$\frac{\partial^2 v}{\partial y_n^2} = \frac{\partial \mu}{\partial y_n} \frac{\partial u}{\partial \nu} + \mu^2 \frac{\partial^2 u}{\partial \nu^2} = \frac{\partial \mu}{\partial y_n} c_1 + \mu^2 c_2$$

$$\frac{\partial^3 v}{\partial y_n^3} = \frac{\partial^2 \mu}{\partial y_n^2} \frac{\partial u}{\partial \nu} + 3\mu \frac{\partial \mu}{\partial y_n} \frac{\partial^2 u}{\partial \nu^2} + \mu^3 \frac{\partial^3 u}{\partial \nu^3}$$

$$= \frac{\partial^2 \mu}{\partial y_n^2} c_1 + 3\mu \frac{\partial \mu}{\partial y_n} c_2 + \mu^3 c_3.$$

Inductively, we see that the transformed functions $\tilde{c}_j$ are obtained in an analytic way from the functions c_i, $i = 0, \ldots, j$.

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. $\tilde{a}_{(0,\ldots,0,k)} \neq 0$.

Compute using the chain rule (remind $\nabla \Psi_n$ is a vector, the gradient of the last coordinate function of Ψ, and it is equal to ν, up to the non-zero multiple μ^{-1})

$$\partial_\alpha u = \frac{\partial^k v}{\partial y_n^k} (\nabla \Psi_n)^\alpha + \text{lot of terms of lower order in } \frac{\partial}{\partial y_n}$$

$$= \mu^{-k} \frac{\partial^k v}{\partial y_n^k} \nu^\alpha + \cdots .$$

647

Substitute into (1)

$$\sum_{|\alpha|=k} a_\alpha \partial_\alpha u = \mu^{-k} \sum_{|\alpha|} a_\alpha \nu^\alpha \frac{\partial^k v}{\partial y_n{}^k} + \cdots .$$

We have computed

$$\tilde{a}_{(0,\ldots,0,k)} = \mu^{-k} \sum_{|\alpha|=k} a_\alpha \nu^\alpha$$

which is non-zero if and only if the orginial 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 Γ 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

$$(1) \qquad \frac{\partial^k}{\partial x_n{}^k} u = \sum_{|\alpha|=k, \alpha_n \neq k} a_\alpha \partial_\alpha u + b$$

and Γ is given by $x_n = 0$ with prescribed normal derivatives $c_0, \ldots, c_{k-1}$. 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 Γ. 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 $\tilde{c}_0, \ldots, \tilde{c}_{\ell-1}$, then we may subtract $g(x_1, \ldots, x_n) = \frac{1}{\ell!} x_n^\ell \tilde{c}_\ell(x_1, \ldots, x_{n-1})$ which kills the next one.

The final reduction step is to introduce new functions for all components in the vector $(u, \nabla u, \ldots, \nabla^{k-1} u)$. Write $v_1, \ldots, v_N$ 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 $v = (v_0, \ldots, v_N)$.

$$(2) \quad \frac{\partial}{\partial x_n} v_s = \sum_{0 \leq r \leq N} \sum_{i=1}^{n-1} a_{sri} \frac{\partial}{\partial x_i} v_r + b_s, \ s = 0, \ldots, N,$$

where all the coefficients a_{sri}, b, are functions in $x_1, \ldots, x_{n-1}, v_0, \ldots, v_N$ 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

$$\tfrac{\partial}{\partial t} t = 1, \quad \tfrac{\partial}{\partial t} u = u_t, \quad \tfrac{\partial}{\partial t} u_x = \tfrac{\partial}{\partial x} u_t,$$

$$\tfrac{\partial}{\partial t} u_t = a_{xt} \tfrac{\partial}{\partial x} u_t + a_{xx} \tfrac{\partial}{\partial x} u_x + b$$

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 Γ 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 coeficients as in the ODE case can be chosen. First, the analycity 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 α.

In particular, all the coefficients can be majorized by the function

$$h(x_1, \ldots, x_{n-1}, v_0, \ldots, v_N) = \frac{Cr}{r - \sum_{j=1}^{n-1} x_j - \sum_{s=0}^{N} v_s}.$$

Now, the majorizing system for the vector $(V_0, \ldots, V_N)$ is

$$\frac{\partial V_s}{\partial x_n} = \sum_{0 \leq r \leq N} \sum_{i=1}^{n-1} h \frac{\partial}{\partial x_i} V_r + h, \ s = 0, \ldots, N.$$

Since the coefficients are completely symmetric in the variables, let us expect the solution in the form

$$V_0 = \cdots = V_N = W(x_1 + \cdots + x_{n-1}, x_n),$$

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.[5]

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

$$(1) \qquad L = \sum_{1 \le i,j \le n} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{1 \le i \le n} a_i \frac{\partial}{\partial x_i} + a$$

written in coordinates $(x_1, \ldots, x_n)$. 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 $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n})$ for the gradient operator, dot for the standard scalar product of vectors, $D^1\Phi$ for the Jacobi matrix of Φ in the coordinates y and x, $\frac{\partial \Phi}{\partial x_i}$ for the ith column of $D^1\Phi$.

$$\frac{\partial}{\partial x_i} = \nabla \cdot \frac{\partial \Phi}{\partial x_i} = \sum_k \frac{\partial \Phi_k}{\partial x_i} \frac{\partial}{\partial y_k}$$

$$\frac{\partial}{\partial x_i \partial x_\ell} = \sum_{k,\ell} \frac{\partial \Phi_k}{\partial x_i} \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial^2}{\partial y_k \partial y_\ell} + \sum_k \frac{\Phi_k}{\partial x_i \partial x_j} \frac{\partial}{\partial y_k}.$$

Thus, the operator (1) transforms into

$$\tilde{L} = \sum_{1 \le i,j,k,\ell \le n} a_{ij} \frac{\partial \Phi_k}{\partial x_i} \frac{\partial \Phi_\ell}{\partial x_j} \frac{\partial^2}{\partial y_k \partial y_\ell} + \sum_{1 \le i,k \le n} a_i \frac{\partial \Phi_k}{\partial x_i} \frac{\partial}{\partial y_k} + a.$$

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 singature, i.e. the number of positive and negative entries in the diagonalized matrix, cf. 4.2.6. This is transfered 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

[5]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 Φ. This will always allow us to transform the quadratic form to its canonical form over the entire domain Ω.

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 genuin mapping Φ 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 Γ will have a locally converning 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) $$L = \frac{\partial^2}{\partial t^2} - c^2 \Delta,$$

where Δ is the Laplace operator $\Delta = \frac{\partial^2}{\partial x_1{}^2} + \cdots + \frac{\partial^2}{\partial x_n{}^2}$, $c^2 > 0$ 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

$$\varphi(x) = f(x) + g(x), \quad \psi(x) = -cf'(x) + cg'(x),$$

651

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 substract 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) $u(x, y) = \dfrac{1}{2}\big(\varphi(x - ct) + \varphi(x + ct)\big) + \dfrac{1}{2c} \displaystyle\int_{x-ct}^{x+ct} \psi(y)\, dy.$

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 α is the separation constant (usually we consider either α^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) $$L = \frac{\partial}{\partial t} - \kappa \Delta,$$

$\kappa > 0$, 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 κ (assumed to be constant here) in time. First of all, let us notice that the usual boundary value presription of the state at time $t = 0$ is not matching the assumption of the Cauchy-Kovalevskaya theroem. 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 inicial 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 accross 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, ℓ. 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 ℓ, $c_{0,2\ell+1} = 0$, $c_{0,2\ell} = (-1)^\ell (2n)!$. 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.).

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 assuptions 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

$$\tilde{u}_t = -\kappa \xi^2 \tilde{u}, \qquad \tilde{u}(\xi, 0) = \tilde{\varphi}.$$

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, . 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 $f(\xi) = \mathrm{e}^{-\kappa t \xi^2}$. But Fourier images $\mathcal{F}(f)$ of Gaussians f are again Gaussians, up to constant, see **??**,

$$\mathcal{F}(\mathrm{e}^{-ax^2})(\xi) = \tfrac{1}{\sqrt{2a}} \, \mathrm{e}^{-\frac{\xi^2}{4a}},$$

with any real constant $a > 0$. Thus, we can write for $t > 0$

$$\tilde{u}(\xi, t) = \mathcal{F}(\varphi)\mathcal{F}\left(\tfrac{1}{\sqrt{2\kappa t}} \, \mathrm{e}^{-\frac{x^2}{4\kappa t}} \right).$$

Finally, we obtain u as the convolution of the initial condition with the so called *heat kernel* function

$$u(x, t) = \frac{1}{2\sqrt{\pi \kappa t}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-y)^2}{4\kappa t}} \, \varphi(y) \, dy.$$

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 $t \to 0_+$ is exactly the function φ, 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 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

$$(1) \quad u_t = \kappa u_{xx}, \; u(x, 0) = 0, \; u(0, t) = 0, \; \frac{\partial}{\partial x} u(d, t) = \rho,$$

where ρ is a constant representing the effect of the heating.

The idea is to exploit the general solutions

$$u(x, t) = (A \cos \alpha x + B \sin \alpha x) \, \mathrm{e}^{\alpha^2 \kappa t}$$

from 9.2.6 with free parameters α, 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 wit the exponentials. This approach is often called the *Fourier method*.

The condition $u(0, t) = 0$ suggests to restrict ourselves to $A = 0$. Then, $u_x(x, t) = B\alpha \cos(\alpha x) \, \mathrm{e}^{-\alpha^2 \kappa t}$. 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

$$u_x(x, t) = B\alpha \cos(\alpha x)\, \mathrm{e}^{-\alpha^2 \kappa t} = 0,$$

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:

$$\rho x - 8\rho d \frac{1}{\pi^2} \sum_{k \geq 0} \frac{(-1)^k}{(2k+1)^2} \sin\left(\frac{(2k+1)\pi x}{2d}\right) \mathrm{e}^{-\kappa \frac{(2k+1)^2 \pi^2 t}{4d^2}}.$$

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 $t \to \infty$, 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 ρ.

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 inpendent 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 Δ 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) $\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + \frac{1}{r}\frac{\partial}{\partial r} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2}.$

The reader should perform the tedious but straightforward computation. Similarly,

(2) $\Delta = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + \frac{\partial^2}{\partial z^2},$

(3) $\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2 \sin^2\psi}\frac{\partial^2}{\partial \varphi^2} + \frac{1}{r^2 \sin\psi}\frac{\partial}{\partial \psi}\left(\sin\psi\frac{\partial}{\partial \psi}\right)$

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, $\Omega = \{x^2 + y^y \leq 1\} \subset \mathbb{R}^2$ 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 α^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 α, including $\alpha = 0$ when Φ 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

$$u(r, \varphi) = C_0 \ln r + D_0 +$$

(4)
$$\sum_{n=1}^\infty (A_n \cos n\varphi + B_n \sin n\varphi)(C_n r^n + D_n r^{-n}).$$

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 distorsion 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 β^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 β 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 Ω, and we know some solution w, i.e. $Lw = f$ (not taking care of the boundary conditions), than we should find a solution v to the homogenenous 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*

$$(1) \qquad Q(x,t) = \frac{1}{\sqrt{4\pi\kappa t}}\, e^{-\frac{x^2}{4\kappa t}}.$$

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 neigborhood 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 $(t - s) \to 0_+$. This leads to

$$\lim_{s \to t_-} \int_{-\infty}^{\infty} Q(x - y, t - s) f(y,s)\, dy = f(x,s).$$

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

$$(3) \qquad \begin{aligned} u(x,t) = &\int_{-\infty}^{\infty} Q(x - y, t)\varphi(y)\, dy + \\ &\int_0^t \left(\int_{-\infty}^{\infty} Q(x - y, t - s) f(y,s)\, dy \right) ds. \end{aligned}$$

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.

$$(4) \qquad Q(x,t) = \frac{1}{\sqrt{(4\pi\kappa t)^n}}\, e^{-\frac{\|x\|^2}{4\kappa t}}.$$

Then taking the n-dimensional (iterated) convolution of Q with the boundary condition φ on the hyperplane $t = 0$ provides the solution candidate

$$(5) \qquad u(x,t) = \int_{\mathbb{R}^n} Q(x - y, t)\varphi(y)\, dy_1 \ldots dy_n.$$

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 allso 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 everywehere expcept 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 => GREEN'S FUNCTION => SOLUTION TO POISSON SIMILARLY TO THE DIFFUSION.

3. Remarks on Variational Calculus

Many practical problems look for minima or maxima of real functions $\mathcal{J} : \mathcal{S} \to \mathbb{R}$ 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 $g(x) = (g_{ij}(x))$, $1 \le i, j \le 2$, 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)

$$(1) \qquad \mathcal{J}(u) = \int_{t_1}^{t_2} \sqrt{g(u(t))(\dot{u}(t))} \, dt$$

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

$$(2) \qquad \mathcal{J}(u) = \int_{t_1}^{t_2} \sqrt{1 + \dot{v}(t)^2} \, dt.$$

Quite certainly we all believe that the mimimum 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 everywehere. 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 α on $\mathbb{R}^2$ and consider $g(x) = \alpha(x)^2 \, \mathrm{id}_{\mathbb{R}^2}$, i.e., the Euclidean size of vectors scaled by $\alpha(x) > 0$ at each point $x \in \mathbb{R}^2$, we obtain

$$(3) \qquad \mathcal{J}(u) = \int_{t_1}^{t_2} \alpha(t, v(t)) \sqrt{1 + \dot{v}(t)^2} \, dt.$$

We can imagine the speed $1/\alpha$ of a moving particle (or light) in the plane depends of the values of α (the smaller is α, the

bigger is the speed) and our problem will be to find the short-est 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 run-ning. If we believe that the straight lines are the minimizers for constant α, then it is clear that we have to find the optimal point $P = (a, p)$ on the bank of the lake where we start swim-ming. 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 φ is the angle betwen our running track and the nor-mal to the boundary of V, while ψ is the angle between our swimming track and the normal to the boundary (draw a pic-ture yourself!). Thus we have recovered the famous Snell law of light diffraction saying that the proportion of the sine val-ues 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 $\mathcal{C}^1(\Omega)$ of all differentiable mappings $u : \Omega \to \mathbb{R}^m$, a C^2 function $F = F(x, y, p) : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{nm}$ and set the *functional*

$$(1) \qquad \mathcal{J}(u) = \int_\Omega F(x, u(x), D^1 u(x))\omega_{\mathbb{R}^n},$$

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 correspond-ing *minimizer* u with prescribed boundary values u on the boundary $\partial\Omega$ (and perhaps some further conditions restrict-ing 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 $F = F(t, y, p) : \mathbb{R}^3 \to \mathbb{R}$,

$$(2) \qquad \mathcal{J}(u) = \int_\Omega F(t, u(t), \dot{u}(t))\, dt.$$

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, picewise differentiable means the one-side derivatives exist at all points).

A maybe shocking example is the following functional:

$$(3) \qquad \mathcal{J}(u) = \int_0^1 (\dot{u}(t)^2 - 1)^2 \, dt$$

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 ± 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:

$$(1) \qquad F(t, y, p) = y^r \sqrt{1 + p^2}$$

$r > 0$, 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 δ's and consider

$$\mathcal{J}(u + \delta v) = \int_{t_1}^{t_2} F(t, u(t) + \delta v(t), \dot{u}(t) + \delta \dot{v}(t)).$$

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)

$$(2)\ \ 0 = \int_{t_1}^{t_2} F_y(t, u(t), \dot{u}(t))v(t) + F_p(t, u(t), \dot{u}(t))\dot{v}(t) \, dt.$$

Integrating the second term in (2) per partes immediately yields (remember $v(t_1) = v(t_2) = 0$)

$$0 = \int_{t_1}^{t_2} \big(F_y(t, u(t), \dot{u}(t))v(t) - \frac{d}{dt}F_p(t, u(t), \dot{u}(t))\big)v(t) \, dt.$$

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)

$$(3) \qquad \frac{d}{dt} F_p(t, u(t), \dot{u}(t)) = F_y(t, u(t), \dot{u}(t)).$$

An equivalent form of this equation for $\dot{u}(t) \neq 0$ is (we omit the arguments t of u and $\dot{u}$)

$$(4) \qquad F_t(t, u, \dot{u}) = \frac{d}{dt}\big(F(t, u, \dot{u}) - \dot{u} F_p(t, u, \dot{u})\big).$$

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

$$(5) \qquad\qquad u = C(1 + \dot{u}^2)^{1/2r}$$

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 τ (which is actually the slope of the tangent to the solution graph):

$$(6) \qquad t = t_0 + \frac{C}{r} \int_0^\tau (\cos s^{-1/r}) ds \quad u = C(\cos \tau)^{-1/r}.$$

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!)

$$(7) \qquad\qquad t = t_0 - C \sin \tau, \quad u = C \cos \tau,$$

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

$$(8) \qquad t = t_0 - \frac{C}{2}(2\tau + \sin 2\tau), \quad \frac{C}{2}(1 + \cos 2\tau)$$

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[6]

$$(9) \qquad u(t) = C \cosh \frac{t - t_0}{C}.$$

9.3.4. Critical points of functionals. Now we shall develope a bit of theory verifying that the steps done in the previous examples realy 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 $\mathcal{F} : \mathcal{S} \to \mathbb{R}$ called (real) *functionals*.

Let $\mathcal{S}$ be a vector space equipped with a norm $\| \ \|$. A continuous linear mapping $L : \mathcal{S} \to \mathbb{R}$ is called a *continuous linear functional*.

A functional $\mathcal{F} : \mathcal{S} \to \mathbb{R}$ 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

$$(1) \qquad \lim_{v \to 0} \frac{\mathcal{F}(u + v) - \mathcal{F}(u) - L(v)}{\|v\|} = 0.$$

---

[6] 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.[7] Indeed, if (1) holds true, than for each fixed $v \in \mathcal{S}$

$$(2) \quad \delta\mathcal{F}(u)(v) = \lim_{t \to 0} \frac{\mathcal{F}(u + tv) - \mathcal{F}(u)}{t} = \frac{d}{dt}_{|0} \mathcal{F}(u+tv)$$

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

$$(3) \quad \delta^2\mathcal{F}(u)(v, w) = \lim_{t \to 0} \frac{\delta\mathcal{F}(u + tv)(w) - \delta\mathcal{F}(u)(w)}{t}.$$

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, $\delta^2\mathcal{F}(u)(w, w) = \frac{d^2}{dt^2}_{|0}\mathcal{F}(u+tw)$, if the second variation exists. We may summarize:

Theorem. *Let $\mathcal{F} : \mathcal{S} \to \mathbb{R}$ 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 $-\mathcal{F}$). $\square$

---

[7]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 $0 < c \leq 1$

$$\mathcal{F}(z) - \mathcal{F}(u) = f(1) - f(0) = f'(0) + \frac{1}{2}f''(c)$$

$$= \frac{1}{2}\delta^2\mathcal{F}(u + cw)(w, w) \geq 0.$$

$\square$

Remark. Actually, the condition from the collolary 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 $\delta^2\mathcal{F}(u)(w, w) \geq C\|w\|$ 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)

$$(1) \qquad \mathcal{J}(u) = \int_{t_1}^{t_2} F(t, u, \dot{u})\, dt$$

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.

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:*

$$(2) \quad \delta \mathcal{J}(u)(v) = \int_{t_1}^{t_2} \left(F_y(t, u, \dot{u})v + F_p(t, u, \dot{u})\dot{v} \right) dt$$

$$(3) \quad \delta^2 \mathcal{J}(u)(v, v) = \int_{t_1}^{t_2} \left(F_{yy}(t, u, \dot{u})v^2 + \right.$$
$$\left. 2F_{yp}(t, u, \dot{u})v\dot{v} + F_{pp}(t, u, \dot{u})\dot{v}^2 \right) dt.$$

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. □

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 6.1.6. 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 φ, 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). □

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) $$F_y(t, u, \dot{u}) - \frac{d}{dt} F_p(t, u, \dot{u}) = 0.$$

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)
$$0 = \int_{t_1}^{t_2} \big(F_y(t, u, \dot{u})v + F_p(t, u, \dot{u})\dot{v} \big) \, dt$$
$$= \int_{t_1}^{t_2} \big(F_y(t, u, \dot{u}) - \frac{d}{dt} F_p(t, u, \dot{u}) \big) v \, dt$$
$$+ F_p(t, u, \dot{u})v_{|t=t_2} - F_p(t, u, \dot{u})v_{|t=t_1}.$$

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. □

9.3.7. Remarks. We made our life comfortable by taking very strong differentiablity 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 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 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 $\tau = t_0 + \frac{t}{\alpha}$)

$$
\begin{aligned}
\delta^2 \mathcal{J}(u)(\varphi, \varphi) &= \int_{t_1}^{t_2} \big(F_{yy}(t, u, \dot{u})v^2 + 2F_{yp}(t, u, \dot{u})v\dot{v} \\
&\qquad\qquad + F_{pp}(t, u, \dot{u})\dot{v}^2 \big)\, dt \\
&\leq \int_{t_0-\alpha s}^{t_0+\alpha s} (C\alpha^2 + 2C\alpha|\dot{v}|)\, dt - \frac{1}{2}\mu \int_{t_0-\alpha s}^{t_0+\alpha s} \dot{v}^2\, dt \\
&= 2Cs\alpha^3 + 2C\alpha^2 \int_{t_0-s}^{t_0+s} |\dot{\varphi}|\, d\tau - \frac{1}{2}\mu\alpha \int_{t_0-s}^{t_0+s} \dot{\varphi}^2\, d\tau \\
&= 2Cs\alpha^3 + 4Cs\alpha^2 - \frac{1}{2}\mu\Big(\int_{t_0-s}^{t_0+s} \dot{\varphi}^2\, d\tau\Big)\alpha.
\end{aligned}
$$

The integral on the right-hand side is strictly positive and, thus, the entire expression is negative if α is small enough. This is a contradiction and the proof is complete. $\qquad\square$

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 straigtforward consequence of the general equation 9.3.6(1), whose equivalent form we saw already in 9.3.3(4)

Case 1. If the Lagrangian is $F(t, y)$, i.e., does not depend on the derivatives, then the Euler-Lagrange equation says

$$(1) \qquad\qquad F_y(t, u) = 0$$

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 Langrangian is $F(t, p)$, then the Euler-Lagrange equation is

$$(2) \qquad\qquad \frac{d}{dt}F_p(t, \dot{u}) = 0$$

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 Langrangian is $F(u, \dot{u})$, then there is a consequence of the Euler-Lagrange equation (for $\dot{u} \neq 0$)

$$(3) \qquad\qquad \frac{d}{dt}\big(F(u, \dot{u}) - \dot{u}F_p(u, \dot{u})\big) = 0$$

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 $f : \Omega \to \mathbb{C}$ is called a *complex function of a single complex variable*. Working with complex numbers, we may repeat the definition of derivative:

<div align="center">COMPLEX DERIVATIVE</div>

We say that a complex function $f : \Omega \to \mathbb{C}$ has the *complex derivative $f'(a)$* at a point $a \in \Omega$, if the complex limit

$$f'(a) = \lim_{z \to a} \frac{f(z) - f(a)}{z - a} \in \mathbb{C}$$

exists. We say that f is *differentiable* in the complex sense, or *holomorphic*, on Ω, 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 $(z+h)^n - z^n = h\big(nz^{n-1} + \frac{1}{2}n(n-1)z^{n-2}h + \cdots + h^{n-1}\big)$ and thus for all $z \in \mathbb{C}$ we obtain the limit

$$(z^n)' = \lim_{h \to 0} \frac{(z+h)^n - z^n}{h} = nz^{n-1}.$$

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 $f : \Omega \to \mathbb{C}$ is called *analytic* in the region Ω 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) $$f(z) = \sum_{n=0}^{\infty} c_n (z - a)^n.$$

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 an any real multivariate calculus) was promised already in i.

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.[8]

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 $|z - a| > R$;*
- *$\frac{1}{R} = \limsup_{n \to \infty} |c_n|^{\frac{1}{n}}$.*

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 $L = \limsup_{n \to \infty} |c_n|^{\frac{1}{n}}$, 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 $n \to \infty$. 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 $\lim n^{\frac{1}{n}} = 1$.

---

[8]Actually the opposite implication is true as well: A holomorhpic function on a domain Ω is analytic on Ω. 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 $g(z) = \sum_{n=1}^{\infty} nc_n(z-a)^{n-1}$ is the term-by-term derivative of $f(z)$. We prove that $f'(z_0) = g(z_0)$, which means

$$\lim_{h \to 0} \frac{f(z_0 + h)f(z_0)}{h} - g(z_0) = 0.$$

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 ε. 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

$$\frac{f(z_0 + h) - f(z_0)}{h} - g(z_0) =$$
$$\left(\frac{S_N(z_0 + h) - S_N(z_0)}{h} - S_N'(z_0) \right) +$$
$$\left(S_N'(z_0) - g(z_0) \right) + \frac{E_N(z_0 + h) - E_N(z_0)}{h}.$$

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 $h \to 0$. 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 $S_N'(z_0) \to g(z_0)$ as $N \to \infty$ (because we know that $g(z)$ is a power series which converges absolutely for z in it's 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

$$E_N(z_0+h)-E_N(z_0) = \sum_{n=N+1}^{\infty} \left(c_n(z_0+h-a)^n - c_n(z_0-a)^n \right)$$

Expanding $(z_0+h-a)^n - (z_0-a)^n = h\big((z_0+h-a)^{n-1} + (z_0+h-a)^{n-2}(z_0-a) + \cdots + (z_0-a)^{n-1}\big)$, we obtain

$$\frac{E_N(z_0 + h) - E_N(z_0)}{h} = \sum_{n=N+1}^{\infty} \big(c_n((z_0 + h - a)^{n-1} +$$
$$(z_0 + h - a)^{n-2}(z_0 - a) + \cdots + (z_0 - a)^{n-1}\big).$$

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:

$$\left| \frac{E_N(z_0 + h) - E_N(z_0)}{h} \right| \leq \sum_{n=N+1}^{\infty} |c_n| n r^{n-1}.$$

The series on the right converges, and furthermore, its value approaches 0 as $N \to \infty$. Indeed, $\sum_{n=N+1}^{\infty} |c_n| n r^{n-1}$ 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 $N \to \infty$. 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:

$$\left| \frac{S_N(z_0 + h) - S_N(z_0)}{h} - S_N'(z_0) \right| +$$

$$\left| S_N'(z_0) - g(z_0) \right| + \left| \frac{E_N(z_0 + h) - E_N(z_0)}{h} \right| \leq \varepsilon \quad \square$$

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.

<small>COROLLARIES ON THE DERIVATIVES OF POWER SERIES</small>

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 $f^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1) \ldots (n-k+1) c_n (z-a)^{n-k}$, 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 $c_k = \frac{f^{(k)}(a)}{k!}$. $\qquad\square$

9.4.4. Links to the real calculus. Each complex valued function f on Ω can be viewed as a mapping $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2$. 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 Ω, with $D^1 f$ being the Jacobi matrix of first partial derivatives in the two real coordinates and $\lim_{z \to a} \alpha(z) = 0$.

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 $f(z) = u(x + i\,y) + i\,v(x + i\,y)$ is the coordinate expression of a complex differentiable function f viewed as a differentiable mapping $\mathbb{R}^2 \to \mathbb{R}^2$, $z = x + i\,y$, then clearly

$$\frac{\partial f}{\partial x}(z) = f'(z) \cdot 1, \quad \frac{\partial f}{\partial y}(z) = f'(z) \cdot i.$$

Thus, $\frac{\partial u}{\partial y} + i \frac{\partial v}{\partial y} = i\left(\frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}\right)$ and we have arrived at the sufficient and necessary conditions for $D^1 f(z)$ being complex linear

$$(1) \qquad u_x = v_y, \quad u_y = -v_x.$$

Yet another argument goes as follows: the rank two matrix describing multiplication by a complex number $a + i\,b$ 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 $f = u + i\,v$.

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 $f : \Omega \subset \mathbb{C} = \mathbb{R}^2 \to \mathbb{C}$ 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* γ in the complex plane is a continuous mapping $\gamma : J \subset \mathbb{R} \to \mathbb{C}$ 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[9], if it does not intersect itself and $\gamma(a) = \gamma(b)$.

---

[9]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 γ with any continuous mapping f defined on the target of γ is again continous, 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_γ along a path γ 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 γ of an interval in $\mathbb{R} \subset \mathbb{C}$.

Writing $f = u + i\,v$ and $\gamma(t) = x(t) + i\,y(t)$,

$$f(\gamma)\dot{\gamma} = (u + i\,v)(\dot{x} + i\,\dot{y}) = (u\dot{x} - v\dot{y}) + i(v\dot{x} + u\dot{y}).$$

Now we may check that the complex value I_γ 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

$$(u + i\,v)(dx + i\,dy) = (u\,dx - v\,dy) + i(v\,dx + u\,dy).$$

Thus, I_γ 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 γ. Thus the integral is independent of the parametrization, up to sign.

If γ is a composition $\gamma_2 \circ \gamma_1$ of two paths (we simply concatenate the curves $\gamma_1 : [a, b] \to \mathbb{C}$ and $\gamma_2 : [b, c] \to \mathbb{C}$, $\gamma_2(b) = \gamma_1(b)$), 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_γ depends continuously on the choice of the path γ in the C^0 metric on the functions on the interval J. Thus we may approximate any integral I_γ 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.

For each piecewise differentiable path $\gamma : [a, b] \to \mathbb{C}$ and antiderivative $F(z)$ of the function $f(z)$ defined on a neighborhood of γ,

$$(1) \qquad I_\gamma = \int_\gamma f(z)\, dz = F(\gamma(b)) - F(\gamma(a)).$$

In particular, the value of the integral depends only on the values of F in the endpoints of γ 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 γ, $\gamma(0) = z$, $\gamma(1) = w$,

$$(F - G)(w) - (F - G)(z) = \int_\gamma 0\, dz = 0,$$

and thus $F - G$ is a constant function. $\qquad\square$

As an example, consider the paths $\gamma_r : [0, 2\pi] \to \mathbb{C}$, $\gamma_r(t) = r\, e^{it}$, 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 $n \in \mathbb{Z}$.

$$(2)\qquad \begin{aligned} \int_{\gamma_r} z^n dz &= \int_0^{2\pi} r^n\, e^{nit}\, ir\, e^{it}\, dt \\ &= \begin{cases} \frac{r^{n+1}}{n+1}[e^{(n+1)it}]_0^{2\pi} = 0 & n \neq -1 \\ ir^0 \int_0^{2\pi} e^0\, dt = 2\pi i & n = -1. \end{cases} \end{aligned}$$

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 Ω is called *simply connected* if every simple closed continuous path in Ω can be shrunk continuously into a point without leaving Ω.

Theorem. *Let $f : \Omega \to \mathbb{C}$ be analytic on a simply connected domain Ω and γ be a closed piecewise differentiable path in Ω. 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 Ω is contained in one such disc. Then we break the closed path $\gamma : [0, 1] \to \Omega$ into the intervals

$0 < t_1 < t_2 < \ldots < t_m = 1$ on which γ is differentiable, and

$$\int_\gamma f(z)\,dz = \sum_{j=0}^{m-1} \big(F(\gamma(t_{j+1})) - F(\gamma(t_j))\big)$$
$$= F(\gamma(1)) - F(\gamma(0)) = 0$$

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 Ω, then $\int_{\partial T} f(z)\,dz = 0$, where ∂T is the oriented boundary of T.

Next, without loss of generality, the path γ 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 γ_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 γ_n can be triangulated into closed triangles T_j so that all T_j lie in Ω with their interiors. Actually, here we need the assumption that Ω is simply connected if we want to fill in all the details. The integral along the path γ_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 γ_n). Finally, possibly refining the polygon γ_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. □

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 holomorhic 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 γ 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

$$\int_{\partial G} f(z)\,dz = \int_G \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + i \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) \right) dx\,dy$$
$$= 2i \int_G \frac{\partial f}{\partial \bar{z}}\,dx\,dy = 0,$$

since $\frac{\partial f}{\partial \bar{z}} = 0$ is equivalent to being holomorphic. □

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 Ω, its antiderivative $F(z)$ exists in that region.*

PROOF. If Ω 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 ζ. Define

$$F(\zeta) = \int_\gamma f(z)\,dz.$$

Choose any other $\tilde{\gamma}$ with the same beginning and end, and prolong the path γ 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 γ.

Next, consider a small h such that the entire oriented segment ν joining ζ and $\zeta + h$ is in Ω. 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

$$\lim_{h \to 0} \frac{F(\zeta + h) - F(\zeta)}{h} = \lim_{h \to 0} \int_0^1 f(\zeta + ht)\,dt = f(\zeta)$$

and thus, $F(\zeta)$ is the requested antiderivative. □

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 Ω should glue together. We shall deal first with the uniqueness issues.

<div align="center">UNIQUENESS THEOREM</div>

Lemma. *Consider an analytic function $f(z)$ in Ω and a sequence of its zeroes $a_n \in \Omega$ which has a limit point $a \in \Omega$. Then $f(z) = 0$ everywhere in Ω.*

PROOF. We start with a simple observation on the non-vanishing of analytic functions:

Claim. *Let $f(z) \neq 0$ be analytic in Ω, 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 nontrivial disc centered at a. Assume $f(w) \neq 0$, $w \in \Omega$, and choose a path γ 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 $c_k = \frac{1}{k!} f^{(k)}(a)$, $k = 0, 1, \ldots$.

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* Φ 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 Φ 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 Ψ is called an *analytic extension* of Φ along the chain $\Phi_0 = \Phi, \Phi_1, \ldots, \Phi_n = \Psi$ if Φ_{j+1} is an immediate analytic extension of Φ_j, $j = 0, 1, \ldots, n-1$.

A parametric family $\Phi_t = (D_t, f_t)$ of canonical elements is called an *analytic extension of Φ_0 along a path* $\gamma : [0,1] \to \mathbb{C}$, 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 Φ_t are $a_t = \gamma(t)$; and (ii) for all $\tau \in [0,1]$ there exist open intervals $U_\tau \subset [0,1]$ containing τ such that for all $t \in U_\tau$, $\gamma(t) \in D_\tau$ and Φ_t is an immediate extension of Φ_τ.

For example, the pair

$$(1) \qquad \log z = \left(|z - 1| < 1, \ \sum_{n=0}^{\infty} \frac{(-1)^n}{n} (z-1)^n \right)$$

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 Ψ 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 Φ_1. Then, Φ_2 is also an immediate extension of Φ_0 if $D_0 \cap D_1 \cap D_2 \neq \emptyset$.*

If Φ_t and Ψ_t are two analytic extensions of canonical element $\Phi_0 = \Psi_0$ along the same path $\gamma : [0,1] \to \mathbb{C}$, 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 Φ_{t_0} and Ψ_{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 Φ_{t_0} and Ψ_{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 Φ_t of elements that extend Φ_0 along some $\gamma(t) : [0,1] \to \mathbb{C}$ 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 Φ_t is a re-expansion of Φ_τ, and therefore $R(t) = \infty$ for all t.

On the other hand, if $R(\tau) < \infty$ for some τ, 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 γ is uniformly continuous on $[0,1]$, $R(t)$ is continuous on U_τ, and thus also on $[0,1]$. $\square$

Lemma. *Consider an analytic extension $\{\Phi_t, \ t \in [0,1]\}$ of Φ_0 along a path γ. There exist a finite number of intermediate points $0 = t_0 < t_1 < \ldots < t_n = 1$ such that Φ_1 is obtained by analytic extension along the chain of elements $\Phi_0 = \Phi_{t_0}, \Phi_{t_1}, \ldots, \Phi_{t_n} = \Phi_1$.*

Conversely, if a canonical element Ψ extends Φ along the chain $\Phi = \Phi_0, \Phi_1, \ldots, \Phi_n = \Psi$ and $\gamma : [0,1] \to \mathbb{C}$ is the piece-wise linear path through the centres of $\Phi_0, \Phi_1, \ldots, \Phi_n$, then there is a family $\{\Psi_t\}$ of canonical elements extending $\Psi_0 = \Phi_0$ along γ, 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 $J_{t_1}, \ldots, J_{t_{n-1}}$ for some $t_1 < \ldots < t_{n-1}$, 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 Φ_{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 Φ_0 or Φ_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 Φ_0 or Φ_1 at the centre $\gamma(t)$ determine same canonical element as noticed above. The interval U_τ can be chosen so that $\gamma(t), t \in U_\tau$ entirely lies in either D_0 or D_1. Then Φ_t for $t \in U_\tau$ will be an immediate extension of Φ_τ. $\square$

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 γ_0 and γ_1 with a common beginning and end must coincide.

<div align="center">HOMOTOPIC PATHS</div>

We say that two paths $\gamma_0(t)$ and $\gamma_1(t)$ with common terminal points $\gamma_0(0) = \gamma_1(0) = a$ and $\gamma_0(1) = \gamma_1(b) = b$ are homotopic if there exists a continuous function $\gamma : [0,1] \times [0,1] \to \mathbb{C}$ 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 γ_0 to γ_1.

The following theorem says that homotopic paths lead to the same analytic extensions.

<div align="center">MONODROMY THEOREM</div>

Theorem. *Suppose that a canonical element $\Phi_0 = (D_0, f_0)$ centred at a can be analytically extended along every path γ_s in a homotopic deformation. Then the extensions along each of those paths terminate with the same canonical element Φ_1.*

PROOF. Write Φ_{st}, $s \in [0,1]$, $t \in [0,1]$ the canonical elements of the extension of $\Phi_{00} = \Phi_0$ along $\gamma_s(t)$. Let $R_s(t)$ be the radius of convergence of Φ_{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 Φ_{st} centred at $\gamma(s,t)$ is a re-expansion of Φ_{s_0t} because the centre $\gamma(s,t)$ lies in the disc of convergence of Φ_{s_0t}. Thus, extensions along $\gamma_s(t)$ and $\gamma_{s_0}(t)$ produce the same terminal element Φ_{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 Φ extends along any path $\gamma \subset \Omega$ through a. Then for any $b \in U$, the extension of Φ 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 Φ to every point in Ω generates an analytic function that is represented as a convergent power series in any disc inscribed in Ω.

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 θ from 0 to 2π, the value of the branches swap. See **??** for more observations on root functions.

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 Ω, 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 Ω 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 Ω 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 Ω and all the others are bounded "holes" inside of Ω. We write γ_0 for the oriented exterior boundary, i.e. the boundary of the connected unbounded component of $\mathbb{C} \setminus \Omega$ oriented counter-clockwise,

while $\gamma_1, \ldots, \gamma_n$ form the oriented interior boundary, i.e. the curves are oriented clockwise.

<div style="text-align:center">

CAUCHY INTEGRAL THEOREM

</div>

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 Ω 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 γ_1 and we may choose any two smooth paths μ_1 and μ_2 joining the left most points and the right most points in γ_0 and γ_1, respectively. This way, we split Ω into two simply connected regions Ω^+ (say the upper one) and Ω^- (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 μ_i and μ_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 γ_i closest to γ_0. Choose two cuts μ_1, μ_2 so that one of the two newly created components of Ω is simply connected. Then the other component has one less interior boundary and thus the theorem follows by induction. $\square$

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 Ω, 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 γ. Indeed, the region enclosed between the two choices γ_1, γ_2 is bounded by them, but with opposite orientations. Thus the vanishing of the integral over the boundary means that actually the integrals over γ_1 and γ_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 γ very small):

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 Ω,*

$$f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(\zeta)}{\zeta - z}\, d\zeta.$$

PROOF. Fix $z \in \Omega$ and consider an open disc $D_\rho = \{\zeta \in \mathbb{C};\ |\zeta - z| < \rho\}$ lying inside of Ω.

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_ρ, 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

$$\left| \int_{\partial D_\rho} \frac{f(\zeta) - f(z)}{\zeta - z}\, d\zeta \right| \le \max_{|\zeta - z| = \rho} \frac{2\pi\rho|f(\zeta) - f(z)|}{\rho}$$
$$= 2\pi \max_{|\zeta - z| = \rho} |f(\zeta) - f(z)|.$$

Clearly, the right hand side approaches zero as $\rho \to 0$. Hence,

$$\int_{\partial D_\rho} \frac{f(\zeta) - f(z)}{\zeta - z}\, d\zeta = 0$$

and the formula in the theorem has been verified. $\qquad \square$

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 Ω 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 $f(z)$:

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 Ω,*

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_{\partial\Omega} \frac{f(\zeta)}{(\zeta - z)^{n+1}}\, d\zeta.$$

PROOF. Indeed, z is an independent argument in the smooth integrand, thus we may differentiate the integral, which yields the formula. $\qquad \square$

Apply the Cauchy integral formula to the disc $D_r = \{|z - a| < r\}$ to obtain

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 $d\zeta = ir\,e^{i\theta}\,d\theta$ to obtain

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + r\,e^{i\theta})d\theta.$$

$\square$

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.

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*

$$(1) \qquad f(z) = \sum_{n=-\infty}^{\infty} c_n(z - a)^n,$$

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 ρ.

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

$$J_1 = \frac{1}{2\pi i} \sum_{n=0}^{\infty} c_n (z - a)^n, \quad c_n = \int_{|\zeta - a| = \rho} \frac{f(\zeta)}{(\zeta - a)^{n+1}} d\zeta.$$

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 ρ. $\qquad \square$

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 $\frac{1}{R} = \limsup_{n \to \infty} |c_n|^{\frac{1}{n}}$, 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 $r = \limsup_{n \to \infty} |c_{-n}|^{\frac{1}{n}}$. Thus we have verified:

Proposition. *For any set of coefficients $\{c_n, n \in \mathbb{Z}\}$ set*

$$\frac{1}{R} = \limsup_{n \to \infty} |c_n|^{\frac{1}{n}}, \quad r = \limsup_{n \to \infty} |c_{-n}|^{\frac{1}{n}}.$$

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

$$c_n = \frac{1}{2\pi i} \int_{|z|=\rho} \frac{f(z)}{z^{n+1}} \, dz = \frac{1}{2\pi} \int_0^{2\pi} f(\mathrm{e}^{it}) \, \mathrm{e}^{-int} \, dt.$$

Therefore, c_n represents the n-th Fourier coefficient of $\varphi(t) = f(\mathrm{e}^{it})$ on $t \in [0, 2\pi]$, and the Fourier series of $f(\mathrm{e}^{it})$ converges uniformly to $f(\mathrm{e}^{it})$ 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) $\quad |c_n| = \left| \frac{1}{2\pi i} \int_{|z-a|=\rho} \frac{f(z)}{(z-a)^{n+1}} \, dz \right| \leq \frac{\max\limits_{|z-a|=\rho} |f(z)|}{\rho^n}$

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 $f(z) = c_0$. $\qquad \square$

9.4.22. Isolated singularities. We look at typical examples of 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 $\lim_{z \to 0} f(z) = 1$.

On the other hand $f(z) = \frac{1}{z}$ grows towards infinity, $\lim_{z\to 0} \frac{1}{z} = \infty$, 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 $\lim_{z\to a} f(z) = \infty$ if and only if $\lim_{z\to 0} |f(z)| = \infty$ 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 $\lim_{z\to a} f(z) = b \in \mathbb{C}$;
- a *pole*, if $\lim_{z\to a} f(z) = \infty$;
- an *essential singularity*, if $\lim_{z\to a} f(z)$ 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 Ω.

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 $|f(z)| \le M$ for $0 < |z - a| < \rho$, then by the Cauchy inequalities 9.4.21(1), $|c_{-n}| \le M\varepsilon^n$, $n > 0$, 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 $\lim_{z \to a} f(z) = \infty$. Therefore $g(z) = \frac{1}{f(z)}$ is also analytic in D and $\lim_{z \to a} g(z) = 0$. 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 $\lim_{z \to a} f(z) = \infty$ 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. $\quad\square$

9.4.23. Some consequences. There are several straightforward corollaries of our classification of isolated singularities. In particular, if $\lim_{z \to a} f(z)$ 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 $z_n \to a$ such that $\lim_{n \to \infty} f(z_n) = w$.*

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 $z_n \to a$ such that $\lim_{n \to \infty} f(z_n) = \infty$.

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 $z \to a$. Therefore $z = a$ is an essential singularity for $g(z)$ and, thus, there is a sequence $z_n \to a$ such that $g(z_n) = \infty$, which implies that $\lim_{n \to \infty} f(z_n) = w$. $\quad\square$

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 ∞ 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, $\lim_{w\to\infty} h(w) = \lim_{w\to\infty} g(w) = f(0)$. 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. □

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 $\overline{D} \setminus \{a_1, \ldots, a_n\}$, where $a_j \in D$, $j = 1, \ldots, n$. 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 $D_j = \{|z - a_j| < \rho\}$, $j = 1, \ldots, n$, have pairwise empty intersections and their closures $\overline{D_j}$ belong to D. Then the Cauchy Integral theorem 9.4.15, applied to $D_\rho = D \setminus \bigcup_{j=1}^{n} D_j$ yields
$$0 = \int_{\partial D_\rho} f(z)\,dz = \int_{\partial D} f(z)\,dz - \sum_{j=1}^{n} \int_{\partial D_j} f(z)\,dz$$
$$= \int_{\partial D} f(z)\,dz - \sum_{j=1}^{n} \text{res}_{a_j} f. \qquad □$$

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 $\mathrm{res}_\infty f = -c_{-1}$.

Note that if $f(z)$ is analytic in $\mathbb{C} \setminus \{a_1, \ldots, a_n\}$, then

$$\mathrm{res}_\infty f + \sum_{j=1}^{n} \mathrm{res}_{a_j} f = 0.$$

Indeed, by taking a disc $\{|z| < R\}$ of sufficiently large radius that does not contain singularities on its boundary, we conclude that

$$\frac{1}{2\pi i} \int_{|z|=R} f(z)\,dz = \sum_{j=1}^{n} \mathrm{res}_{a_j} f.$$

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. $M(R) = \max_{|z|=R, \mathrm{Im}\, z \geq 0} |f(z)|$.

JORDAN'S LEMMA

Lemma. *Consider the function $f(z)$ continuous on $\{\mathrm{Im}\, z \geq 0, |z| = R\}$. Then, for each positive real parameter t,*

$$\int_{|z|=R, \mathrm{Im}\, z \geq 0} f(z)\,\mathrm{e}^{itz}\,dz \leq \frac{\pi}{t} M(R).$$

Consequently, if $f(z)$ is continuous on $\{\mathrm{Im}\, z \geq 0, |z| \geq R_0\}$ and $\lim_{R\to\infty} M(R) = 0$, then

$$\lim_{R\to\infty} \int_{|z|=R, \mathrm{Im}\, z \geq 0} f(z)\,\mathrm{e}^{itz}\,dz = 0.$$

PROOF. We estimate the integral from the lemma:

$$\left| \int_0^\pi f(R\,\mathrm{e}^{i\theta})\,\mathrm{e}^{-tR\sin\theta + itR\cos\theta}\, iR\,\mathrm{e}^{i\theta}\,d\theta \right|$$

$$\leq R\,M(R) \int_0^\pi \mathrm{e}^{-tR\sin\theta}\,d\theta.$$

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

$$R\,M(R) \int_0^\pi \mathrm{e}^{-tR\sin\theta}\,d\theta = 2R\,M(R) \int_0^{\frac{\pi}{2}} \mathrm{e}^{-tR\sin\theta}\,d\theta$$

$$\leq 2R\,M(R) \int_0^{\frac{\pi}{2}} \mathrm{e}^{\frac{-2tR\theta}{\pi}}\,d\theta = \frac{\pi}{t}M(R) \int_0^{Rt} \mathrm{e}^{-\tau}\,d\tau$$

$$= \frac{\pi}{t}M(R)(1 - \mathrm{e}^{-Rt}) \leq \frac{\pi}{t}M(R).$$

The consequence for $R \to \infty$ is obvious. $\square$

Typically, the Jordan lemma is used to compute improper integrals of real analytic (complex valued) functions $g(x) =$

$f(x)\,\mathrm{e}^{itx}$ 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 $\lim_{R\to\infty} M(R) = 0$, then we may compute the real integral

$$\int_{-\infty}^{\infty} g(x)\,dx = \lim_{R\to\infty} \int_{\gamma_R} g(z)\,dz = 2\pi i \sum_k \mathrm{res}_{a_k}\, g(z),$$

where γ_R is the path composed byf 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 Ω 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 holomorhic 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.

Statistics and probability methods

Is statistics a part of mathematics?

– whenever it is so, we need much of mathematics there...!
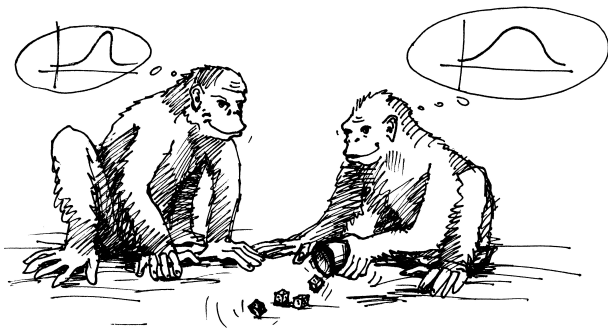
A. Dots, lines, rectangles

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.

$(= N_i/20 = \sum_{j=1}^{i} p_j)$:

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*:
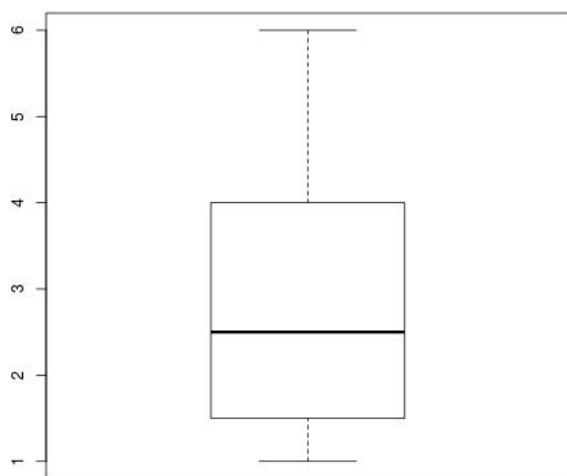
The upper and lower sides of the "box" correspond respectively to the first (lower) and the third (upper) quartile, so its height is equal to the interquartile range. The thick horizontal line is drawn at the median level; the lower and upper horizontal lines correspond respectively to the minimum and 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.

Histogram of x
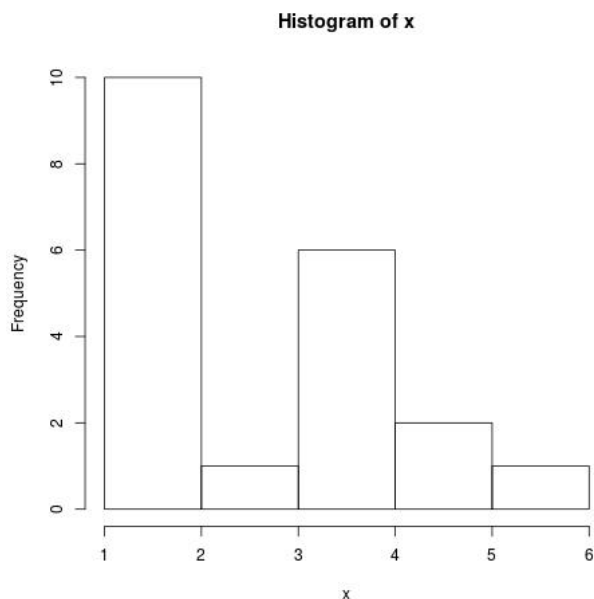
Note that the frequencies of one- and two-member households were merged into a single rectangle. This is used in order to make the data "easier to read" – there exist (various and ambiguous) rules for the merging.

We simply mention this fact without presenting an exact procedure (it is just as anyone likes). $\square$

10.A.2. Given a data set $x = (x_1, x_2, \ldots, x_n)$, 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

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right)^2 = \frac{1}{s_x^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1. \qquad \square$$

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:

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2 \right) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{2\bar{x}}{n} \sum_{i=1}^{n} x_i + \bar{x}^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

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 last 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:

$$\frac{5.1^2 + 4.1^2 + 4 \cdot 1.1^2 + 4 \cdot 0.1^2 + 6 \cdot 0.9^2 + 3 \cdot 1.9^2 + 2.9^2}{1 + 1 + 4 + 4 + 6 + 3 + 1}$$
$$= 3.59.$$

The histogram and box diagram are shown in the following pictures.
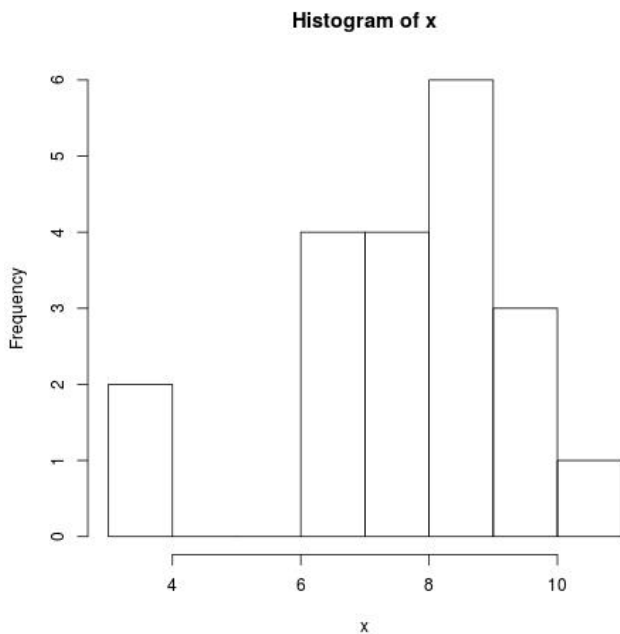
Histogram of x

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,

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* $x_1, x_2, \ldots, x_n$, 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*

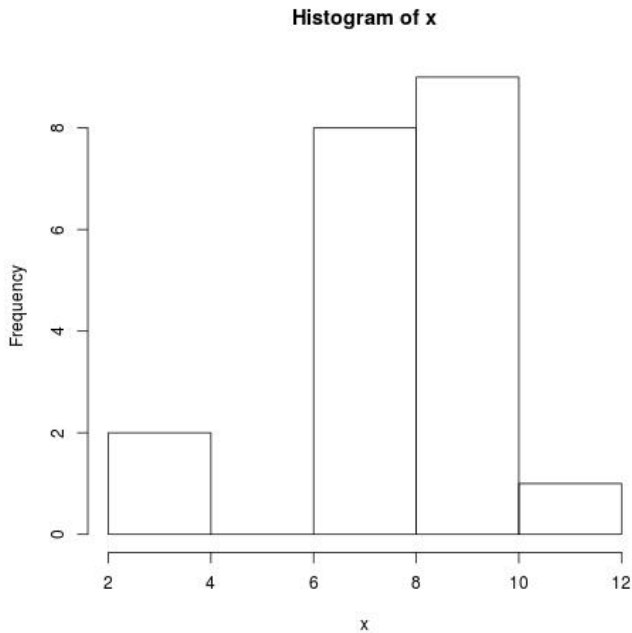$$(1) \qquad\qquad x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

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.

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:

Histogram of x

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:

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, χ^2, respectively).
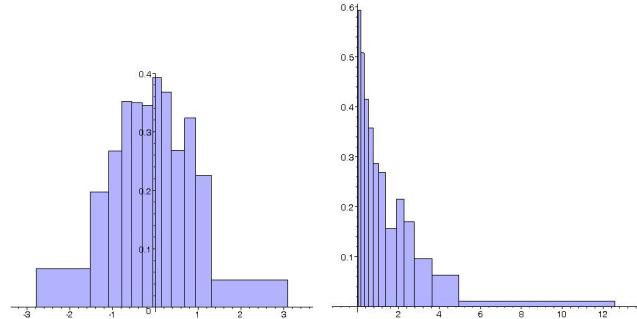
10.1.4. Measures of the position of statistical values. If the magnitude of values around which the collected data values gather are to be expressed, then the concepts of the definition below can be used. There, we work with ratios or interval types of scales.

Consider an (unsorted) data set $(x_1, \ldots, x_n)$ of the values for all examined statistical units and let $n_1, \ldots, n_m$ be the class frequencies of m distinct values $a_1, \ldots, a_m$ 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 useded 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,
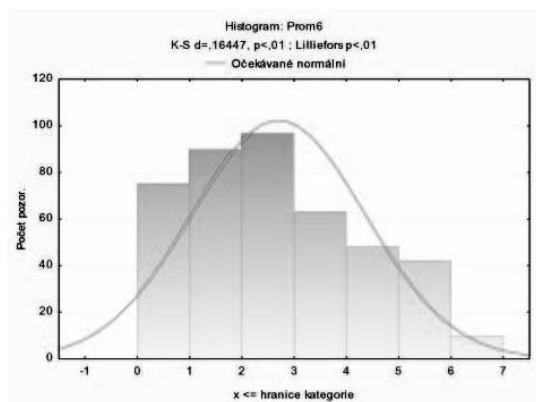$$\overline{(a + b \cdot x)} = \frac{1}{n} \sum_{i=1}^{n} (a + b x_i) = a + b \sum_{i=1}^{n} x_i = a + b \cdot \bar{x}.$$

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 $\tilde{x} = x_{0.5} = 2.5$ 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 $s_x^2 = 2.7$ kg^2, whence it follows that the standard deviation is $s_x = 1.7$ kg, and the coefficient of variation is $V_x = 0,6$. □

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

$$L(p_1, \ldots, p_n, \lambda) = -\sum_{i=1}^{n} p_i \ln p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

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 $i = 1, \ldots, n$. 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 $p_i = \frac{1}{n}$. □

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 8.1.29 (page 547), 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

$$\bar{x}^H \le \bar{x}^G \le \bar{x}.$$

10.1.5. Median, quartile, decile, percentile, … Another way of expressing the position or distribution of the values is to find, for a number α between zero and one, such a value x_α that $100\alpha\%$ of values from the set are at most x_α and the remaining ones are greater than x_α. If such a value is not unique, one can choose the mean of the two nearest possibilities.

The number x_α is called the α–*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_α are the following:

- The *median* (also sample median) is defined by

$$\tilde{x} = x_{0.50} = \begin{cases} x_{((n+1)/2)} & \text{for odd } n \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{for even } n \end{cases},$$

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 α–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:

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:

# 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:
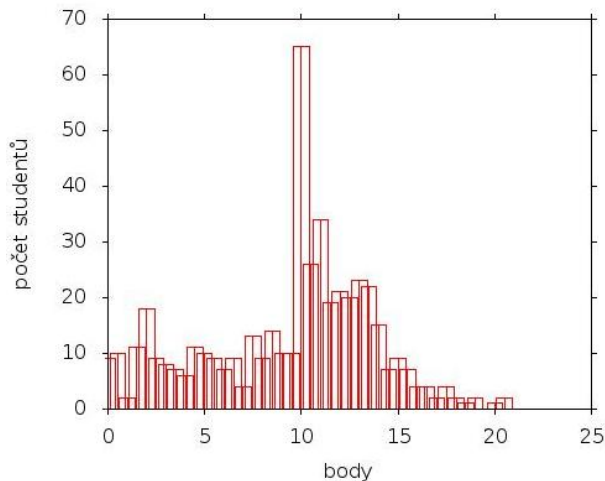
The histogram was obtained from the Information System of Masaryk University. We can see that the data are shown in a somewhat unusual way: individual amounts of points correspond to "double rectangles". It is a matter of taste how to represent the data (it is possible to merge some

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

$$y = (x_1 + c, x_2 + c, \ldots, x_n + c)$$

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

values, thereby decreasing the number of rectangles, or to use thinner rectangles).
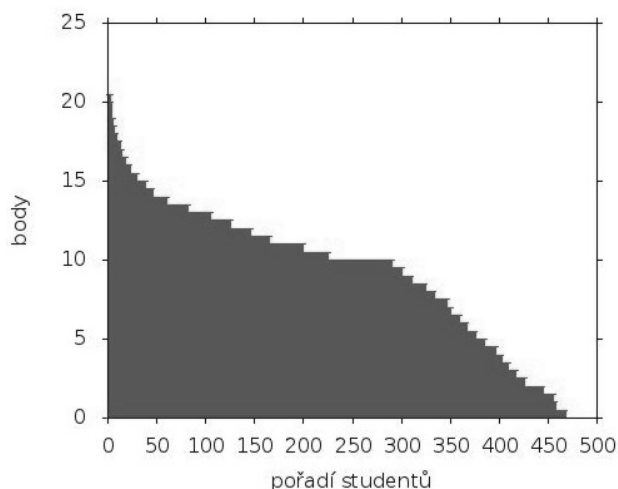
We can notice that the mode of the values is 10, which, accidentally, was also the number of points necessary to pass the course. The mean of the obtained points is 9.48.

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.
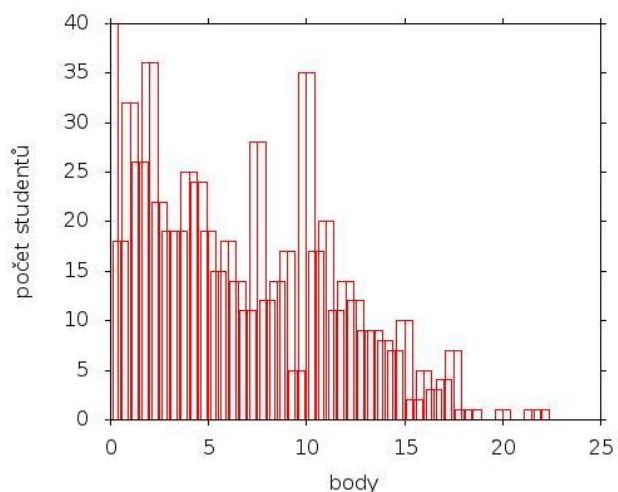
Again, the results can be depicted in an alternative way:

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 β) 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*.
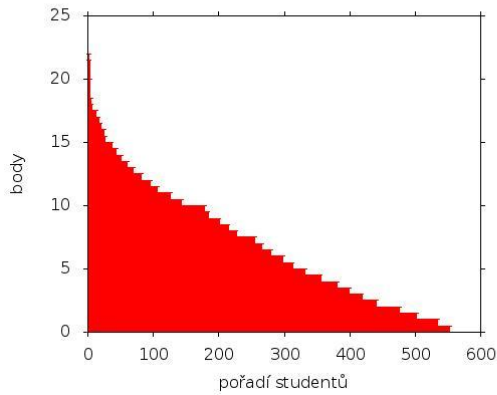
And these are the graphs of amounts of points obtained by those students who continued their studies:
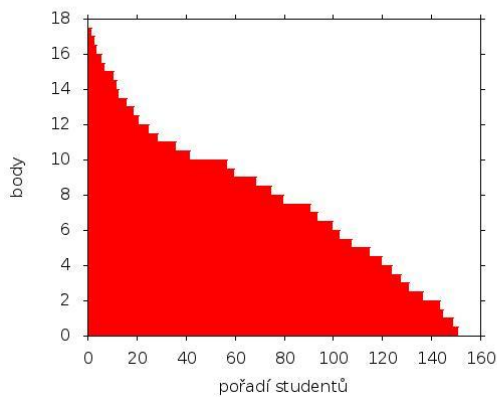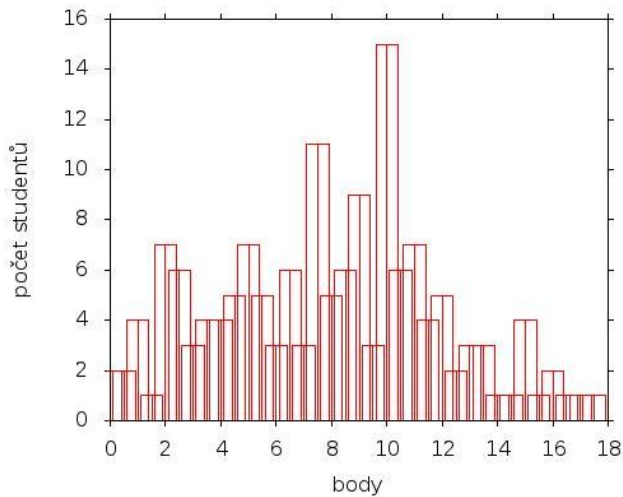

We can see that in the former case, the mode is equal to 0, while in the latter case, it is 10 again. The frequency distribution is close to the one of the MB104 course, which is recommended for the fourth semester.

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.
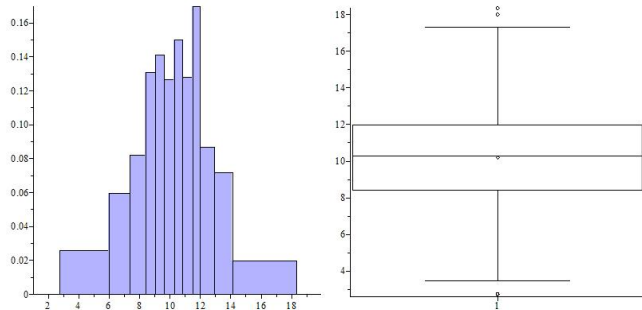
Common displaying tools allow us to view potential dependencies of two data sets. For instance, in the left-hand diagram below, the coordinates are chosen as the values of two independent normal distributions with mean equal to 10 and variance equal to 3. In the right-hand illustration, the first coordinate 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.
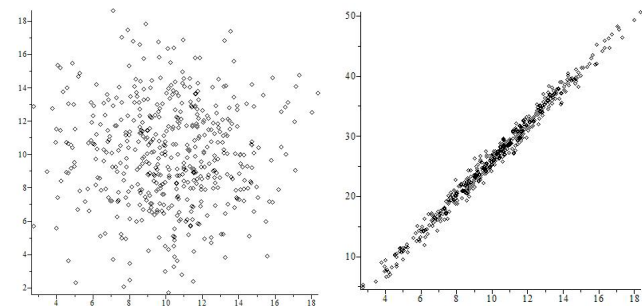
10.1.9. Covariance matrix. Actually, the depencies between several data sets associated to the same statistical units are at the core of our interest in many real world problems. When definining the variance in 10.1.6 above, we employed the euclidean distance, i.e. we evaluated the scalar product of 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.

$\square$

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
  $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, i = 1, \ldots, n,$
- the mean of the j-th component
  $m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, j = 1, \ldots, p,$
- the sample variance of the j-th component
  $s_j = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - m_j)^2, j = 1, \ldots, p,$
- the vector of means $\mathbf{m} = (m_1, \ldots, m_p),$
- 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

Consider two data sets $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, 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 $x^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$, …, $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$, 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 covarinace 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 bout -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 coffeicients 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 λ_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 spaned 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

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 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:

$$\begin{pmatrix} 0.04 & 0.14 & 1.72 \\ 0.14 & 0.49 & 6.02 \\ 1.72 & 6.02 & 73.96 \end{pmatrix}, \quad \begin{pmatrix} 4.840 & 2.64 & 21.12 \\ 2.64 & 1.44 & 11.52 \\ 21.12 & 11.52 & 92.16 \end{pmatrix},$$

$$\begin{pmatrix} 0.641 & 0.640 & 3.521 \\ 0.640 & 0.640 & 3.52 \\ 3.521 & 3.52 & 19.36 \end{pmatrix}, \quad \begin{pmatrix} 0.641 & 0.640 & 2.721 \\ 0.640 & 0.640 & 2.72 \\ 2.721 & 2.72 & 11.56 \end{pmatrix},$$

$$\begin{pmatrix} 0.641 & 0.240 & 8.321 \\ 0.240 & 0.09 & 3.12 \\ 8.32 & 3.12 & 108.16 \end{pmatrix}.$$

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 $(\alpha_1, \ldots, \alpha_k)$ corresponding to the chosen eigenvalue λ, then the values α_j provide the right coefficients in the orthogonal projection $(x^{(1)}, \ldots, x^{(k)}) \mapsto \hat{x} = \alpha_1 x^{(1)} + \cdots + \alpha_k x^{(k)}$.

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)

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_i p_i F(p_i) + \sum_j q_j F(q_j).$$

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:

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 $\mathbf{x}_1 = (1, 1, 2, 2, 1), \ldots$ $\mathbf{x}_{10} = (2, 3, 1, 2, 1)$. The corresponding covariance matrices are:

$$\begin{pmatrix} 1.21 & 1.10 & -0.330 & -0.330 & 0.110 \\ 1.10 & 1. & -0.300 & -0.300 & 0.100 \\ -0.330 & -0.300 & 0.0900 & 0.0900 & -0.0300 \\ -0.330 & -0.300 & 0.0900 & 0.0900 & -0.0300 \\ 0.110 & 0.100 & -0.0300 & -0.0300 & 0.0100 \end{pmatrix}$$

$$\begin{pmatrix} 0.0100 & -0.100 & 0.0701 & -0.0300 & 0.0100 \\ -0.100 & 1. & -0.700 & 0.300 & -0.100 \\ 0.0701 & -0.700 & 0.490 & -0.210 & 0.0701 \\ -0.0300 & 0.300 & -0.210 & 0.0900 & -0.0300 \\ 0.0100 & -0.100 & 0.0701 & -0.0300 & 0.0100 \end{pmatrix}$$

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

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

$$\Big(\sum_j q_j\Big)\Big(\sum_i p_i F(p_i)\Big) + \Big(\sum_i p_i\Big)\Big(\sum_j q_j F(q_j)\Big),$$

leading to

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_{i,j} p_i q_j \big(F(p_i) + F(q_j)\big).$$

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

$$H_X = -\sum_{i=1}^{k} \frac{n_i}{n} \ln\Big(\frac{n_i}{n}\Big),$$

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

$$e^{H_X} = \Big(\big(\tfrac{1}{k}\big)^{-\frac{1}{k}}\Big)^k = k,$$

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.
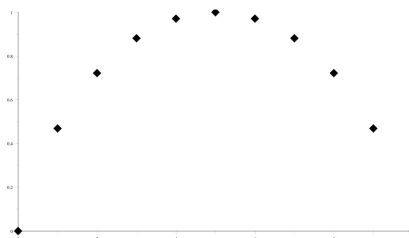
Note that the maximum entropy 1 occurs for the same number of a's and b's, and indeed $2^1 = 2$ as computed above.

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 occurences 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

component is $(1.58; 2.73; 2.13; 2.93; 1.45; 1.93; 4.28; 3.48; 5.26; 3.76)$) 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).

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:

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 Ω 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.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 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

$$10+20+\cdots+10\cdot2^{n-1} = 10\cdot\left(\sum_{i=0}^{n-1} 2^i\right) = 10\cdot\left(\frac{2^n - 1}{2 - 1}\right) = 10\cdot(2^n-1).$$

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

$$\left(1 - \left(\frac{19}{37}\right)^8\right)^{255} \doteq 0.29.$$

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

Therefore, the probability of winning using his strategy is much lower than if he bet everything on red straightaway. □

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 $B_1, B_2, \ldots, B_n$. Using the definition of conditional probability (see 10.2.6), prove that

(1) $$P(A|B) = \sum_{i=1}^{n} P(A|B_i)P(B_i|B)$$

Solution. First, note that the events $A \cap B_1$, $A \cap B_2$,…,$A \cap B_n$ are also disjoint. Therefore, we can write

$$P(A|B_1 \cup \cdots \cup B_n) = \frac{P(A \cap (B_1 \cup \cdots \cup B_n))}{P(B_1 \cup \cdots \cup B_n)} =$$
$$= \frac{P((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n))}{P(B)} =$$
$$= \frac{\sum_{i=1}^{n} P(A \cap B_i)}{P(B_i)} \cdot \frac{P(B_i)}{P(B)} =$$
$$= \sum_{i=1}^{n} P(A|B_i)P(B_i|B).$$

□

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

a) the balls are of different colors;

b) 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

a) 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.

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 Ω with sample points (outcomes) ω_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. σ-fields. Work with a fixed non-empty set Ω, 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 *σ-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.).

σ−ALGEBRAS OF SUBSETS

A collection $\mathcal{A}$ of subsets of the sample space is called a *σ-algebra* or *σ-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

$$\Omega \setminus \cup_{i \in U} A_i^c = \cap_{i \in I} A_i$$

is also in the σ-algebra $\mathcal{A}$.

Altogether, a σ-algebra is a collection of subsets of the sample space which is closed with respect to set differences, countable unions, and countable intersections.

b) 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 used 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

$$
\begin{aligned}
P(B|A) &= P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) \\
&= P(B|A_1) \cdot \frac{P(A_1)}{P(A)} + P(B|A_2) \cdot \frac{P(A_2)}{P(A)} + P(B|A_3) \cdot \frac{P(A_3)}{P(P(A))} = \\
&= 1 \cdot \frac{4}{9} + \frac{2}{3} \cdot \frac{3}{9} + \frac{1}{3}\frac{2}{9} = \frac{20}{27}.
\end{aligned}
$$

$\square$

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, $P(B|A_3) = \frac{8}{12}$

10.2.3. Probability space. Now introduce probability in the mathematical model, recalling the concepts used already in the first chapter.

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:

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

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

and $P(B|A_4) = \frac{9}{12}$. Applying (5), we get that the wanted probability is

$P(B|A) = P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A) =$
$$= \tfrac{7}{12} \cdot \tfrac{1}{6} + \tfrac{8}{12} \cdot \tfrac{1}{3} + \tfrac{9}{12} \cdot \tfrac{1}{2} = \tfrac{25}{36}. \qquad \square$$

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 $P(B|A_i)$, $i = 1, \ldots, 4$. The probability $P(B|A_1)$ can be 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

$P(B|A_1) = P(B_2|A_1) + P(B_3|A_1) + P(B_4|A_1) = \frac{1}{3}\frac{3}{4} + \frac{1}{3}\frac{1}{3} + \frac{1}{3}\frac{1}{4} = \frac{13}{36}$

Similarly,

$$P(B|A_2) = \tfrac{1}{3}\tfrac{1}{2} + \tfrac{1}{3}\tfrac{1}{3} + \tfrac{1}{3}\tfrac{1}{4} = \frac{13}{36},$$
$$P(B|A_3) = \tfrac{1}{3}\tfrac{1}{2} + \tfrac{1}{3}\tfrac{3}{4} + \tfrac{1}{3}\tfrac{1}{4} = \frac{1}{2},$$

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, $\ldots$, again results in a collection of stochastically independent events, and (see 1.4.7, page 23)

$P(A_1 \cup \cdots \cup A_k) = 1 - P(A_1^c \cap \cdots \cap A_k^c) =$
$$= 1 - (1 - P(A_1))\ldots(1 - P(A_k)).$$

Classical finite probability remains the fundamental example of probability, used as the inspiration during creation of the mathematical model. Recall that in this case, Ω is a finite set, the σ-algebra $\mathcal{A}$ is the collection of all subsets of Ω, and the *classical probability* is the probability space $(\Omega, \mathcal{A}, P)$ with probability function $P : \mathcal{A} \to \mathbb{R}$,
$$P(A) = \frac{|A|}{|\Omega|}.$$

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 Ω.

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* $A_1 \subset A_2 \subset \ldots$. *Then,*
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i).$$

Similarly, if $A_1 \supset A_2 \supset A_3 \supset \ldots$, *then*
$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i).$$

PROOF. The considered union $A = \cup_{i=1}^{\infty} A_i$ can be rewritten in terms of mutually exclusive events
$$\tilde{A}_i = A_i \setminus A_{i-1},$$
defined for all $i = 2, 3, \ldots$. Set $\tilde{A}_1 = A_1$. Then,
$$P(A) = P\left(\bigcup_{i=1}^{\infty} \tilde{A}_i\right) = \sum_{i=1}^{\infty} P(\tilde{A}_i) = \lim_{k \to \infty} \sum_{i=1}^{k} P(\tilde{A}_i).$$

$$P(B|A_4) \quad = \tfrac{1}{3}\tfrac{1}{2} + \tfrac{1}{3}\tfrac{3}{4} + \tfrac{1}{3}\tfrac{1}{3} \quad = \frac{19}{36}.$$

Altogether, we get

$$P(B|A) = P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A)$$

$$= \frac{4}{9}\frac{3}{11} + \frac{13}{36}\frac{9}{22} + \frac{1}{2}\frac{2}{11} + \frac{19}{36}\frac{3}{22} = \frac{19}{44}$$

$\square$

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... there are two hits in the target, both of the first shooter,

B... 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:

$$P(B) = \sum_{i=1}^{6} P(B_i) = 169/625.$$

Now, we can compute the conditional probability, using the formula of subsection 10.2.6:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B_5)}{P(B)} = \frac{\frac{64}{625}}{\frac{169}{625}} = \frac{64}{164} \cdot 0.38.$$

$\square$

For the finite sums,

$$\sum_{i=1}^{k} P(\tilde{A}_i) = P(A_1) + \sum_{i=2}^{k} \big(P(A_i) - P(A_{i-1})\big) = P(A_n)$$

by the assumptions $A_i \supseteq A_{i-1}$. 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

$$P(A) = 1 - P(B) = 1 - \lim_{i \to \infty} P(B_i) = \lim_{i \to \infty} \big(1 - P(B_i)\big)$$

which completes the proof. $\square$

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 σ-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 prior 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.

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), $i \in \{0, 1, 2, \ldots, n\}$. 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 $\bigcup_{i=0}^{n}(C \cap$

Lemma. *Let an event B be the union of mutually exclusive events $B_1, B_2, \ldots, B_n$. Then,*

$$(1) \qquad P(A|B) = \sum_{i=1}^{n} P(A|B_i) P(B_i|B)$$

PROOF. The events $A \cap B_1$, $A \cap B_2$, $\ldots$, $A \cap B_n$ are also mutually exclusive. Therefore,

$$P(A|B_1 \cup \cdots \cup B_n) = \frac{P(A \cap (B_1 \cup \cdots \cup B_n))}{P(B_1 \cup \cdots \cup B_n)} =$$

$$= \frac{P((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n))}{P(B)} =$$

$$= \sum_{i=1}^{n} \frac{P(A \cap B_i)}{P(B_i)} \frac{P(B_i)}{P(B)} =$$

$$= \sum_{i=1}^{n} P(A|B_i) P(B_i|B). \qquad \square$$

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:

<div align="center">BAYES' RULES</div>

Theorem. *The probabilities of events A and B satisfy*

$$(1) \qquad P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

$$(2) \qquad P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}.$$

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. $\square$

B_i), we can write

$$P(C|A) = P\left(\bigcup_{i=0}^{n}(C \cap B_i)|A\right) = \sum_{i=0}^{n}\frac{P((C \cap B_i) \cap A)}{P(A)} =$$

$$= \frac{1}{P(A)}\sum_{i=0}^{n}P(C \cap (A \cap B_i)) =$$

$$= \frac{1}{P(A)}\sum_{i=0}^{n}P(A \cap B_i)P(C|A \cap B_i) =$$

$$= \frac{1}{P(A)}\sum_{i=0}^{n}P(B_i)P(A|B_i)P(C|A \cap B_i).$$

We use the law of total probability and substitute for $P(A)$, which leads to

$$(1) \quad P(C|A) = \frac{\sum\limits_{i=0}^{n}P(B_i)P(A|B_i)P(C|A \cap B_i)}{P(A)} =$$

$$= \frac{\sum\limits_{i=0}^{n}P(B_i)P(A|B_i)P(C|A \cap B_i)}{\sum\limits_{i=0}^{n}P(B_i)P(A|B_i)}.$$

This formula is sometimes called the *second Bayes' formula*; it holds in general provided the space Ω 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:

$$\sum_{i=0}^{n}P(B_i)P(A|B_i) = \sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i} \cdot \frac{i}{n} =$$

$$= \sum_{i=1}^{n}\frac{(n-1)!}{(i-1)!(n-i)!}p^i(1-p)^{n-i} =$$

$$= \sum_{i=0}^{n-1}\frac{(n-1)!}{i!(n-i-1)!}p^{i+1}(1-p)^{n-i-1} =$$

$$= p\sum_{i=0}^{n-1}\binom{n-1}{i}p^i(1-p)^{n-1-i} =$$

$$= p(p+(1-p))^{n-1} = p,$$

$$\sum_{i=0}^{n}P(B_i)P(A|B_i)P(C|A \cap B_i) =$$

$$= \sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i} \cdot \frac{i}{n} \cdot \frac{n-i}{n-1} =$$

$$= \sum_{i=1}^{n-1}\frac{(n-2)!}{(i-1)!(n-i-1)!}p^i(1-p)^{n-i} =$$

$$= \sum_{i=0}^{n-2}\frac{(n-2)!}{i!(n-2-i)!}p^{i+1}(1-p)^{n-i-1} =$$

Bayes' rule is sometimes formulated in a somewhat more general form, proved similarly as in (2):

Let the sample space Ω be the union of mutually exclusive events $A_1,\ldots A_n$. Then, for any $i \in \{1,\ldots,n\}$,

$$(3) \quad P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n}P(B|A_k)P(A_k)}.$$

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 10.2.7(2). There is no need to calculate $P(A)$ at all:

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} =$$

$$= \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + 0.5 \cdot 0.8} = \frac{1}{3}.$$

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

$$= p(1-p) \sum_{i=0}^{n-2} \binom{n-2}{i} p^i (1-p)^{n-2-i} =$$

$$= \begin{cases} p(1-p), & n > 1 \\ 0, & n = 1. \end{cases}$$

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

$$P(H) = \frac{1}{3}, \quad P(H^c) = 1 - \frac{1}{3} = \frac{2}{3}.$$

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 $X : \Omega \to \mathbb{R}$ in the sample space Ω 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 σ-algebra:[1]

---

[1]In this connection, we also talk about the σ-algebra of *Borel-measurable sets* on $\mathbb{R}^k$, and then the following definition says that random variables are *Borel-measurable functions*.

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

$$P(C|A \cap H) = 0 = P(C|A^c \cap H^c),$$
$$P(C|A^c \cap H) = 1 = P(C|A \cap H^c).$$

It follows from the second Bayes' formula (1) that

$$P(C|A) =$$
$$= \frac{P(H)P(A|H)P(C|A \cap H) + P(H^c)P(A|H^c)P(C|A \cap H^c)}{P(A)}$$
$$= P(H^c) = \frac{2}{3}$$

and, analogously,

$$P(C|A^c) =$$
$$= \frac{P(H)P(A^c|H)P(C|A^c \cap H) + P(H^c)P(A^c|H^c)P(C|A^c \cap H^c)}{P(A^c)}$$
$$= P(H) = \frac{1}{3}.$$

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 $\frac{1}{2}$. □

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? ○

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?

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 σ-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 $X : \Omega \to \mathbb{R}$. 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 $X : \Omega \to \mathbb{R}$ 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 $X : \Omega \to \mathbb{R}$ 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* $X = (X_1, \ldots, X_k)$ on $(\Omega, \mathcal{A}, P)$ is a k-tuple of random variables $X_i : \Omega \to \mathbb{R}$ defined on the same probability space $(\Omega, \mathcal{A}, P)$.

If intervals $I_1, \ldots, I_k$ 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 $B = I_1 \times \cdots \times I_k$, $P_X(B) = P(X^{-1}(B))$ (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 $X = (X_1, \ldots, X_k)$, write $P(a_1 < X_1 < b_1, \ldots, a_k < X_k < b_k)$ 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.).

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 α ($\alpha \in [0, \pi/2]$) 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 (α, d) is a rectangle $\pi/2 \times l/2$. The favorable events (α, d) (i. e. those for which $l/2 \sin \alpha > d$) correspond to those points in the rectangle which lie under the curve $l/2 \sin \alpha$ (α being the variable of the x-axis). By 6.2.21, the area of the figure is

$$\int_0^{\frac{\pi}{2}} \frac{l}{2} \sin \alpha \, d\alpha = \frac{l}{2}.$$

Thus, the wanted probability is (see **??**)

$$\frac{\frac{l}{2}}{\frac{\pi}{2} \cdot \frac{l}{2}} = \frac{2}{\pi}.$$

$\square$

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

Definition. The *distribution function* or *cumulative distribution function* of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined for all $x \in \mathbb{R}$ by

$$F_X(x) = P(X < x).$$

The *distribution function* of a random vector $(X_1, \ldots, X_k)$ is the function $F_X : \mathbb{R}^k \to \mathbb{R}$ defined for all vectors $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$ by

$$F_X(x) = P(X_1 < x_1, \ldots, X_k < x_k).$$

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.[2]

10.2.12. Theorem. *For every random variable X, its distribution function $F : \mathbb{R} \to [0, 1]$ 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

$$\lim_{x \to \infty} F(x) = 1, \qquad \lim_{x \to -\infty} F(x) = 0;$$

(5) the probability of X taking on the value x is given by

$$P(X = x) = \lim_{y \to x_+} F(y) - F(x).$$

(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,

$$P(A) = \lim_{n \to \infty} P(A_n).$$

----

[2]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.

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.

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

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,

$$P(A) = \lim_{n \to \infty} P(A_n) = P(X \leq x),$$

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 $r_n \to \infty$. Their union is the universal event Ω. In the second case, use the events A_n given by $X < r_n$, for any decreasing sequence $r_n \to -\infty$, 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: $M = \cup_{n=2}^{\infty}$. 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. □

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 σ-algebra of all the Borel sets in $\mathbb{R}$.

In this sense, every function $F : \mathbb{R} \to \mathbb{R}$ 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.

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. □

E. Random variables, density, distribution function

10.E.1. Consider rolling a die. The set of sample points is $\Omega = \{\omega_1, \ldots, \omega_6\}$, where ω_i means that we have rolled number i. Further, consider the σ-field

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}.$$

Find whether the mapping $X : \Omega \to \mathbb{R}$ 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 σ-field. Then, by definition 10.2.10, a random variable is any function $X : \Omega \to \mathbb{R}$ 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 Ω. In each case, the preimage lies in the σ-field $\mathcal{A}$. □

10.E.2. Consider a σ-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 $X : \Omega \to \mathbb{R}$, as general as possible, which is a random variable with respect to $\mathcal{A}$.

Solution. Since the events ω_1, ω_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

If a random variable X assumes only finitely many values $x_1, x_2, \ldots, x_n \in \mathbb{R}$ or countably infinitely many values $x_1, x_2, \ldots$, 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

$$P(x \le X < x + dx) = f(x)dx.$$

This means that whenever $-\infty \le a \le b \le \infty$, it is required that

$$P(a \le X < b) = \int_a^b f(x)dx.$$

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 Ω on which X would be defined as a function. X is represented just by its density or distribution function.

is $\{\omega_1, \omega_2, \omega_4, \omega_5\} \in \mathcal{A}$, which is okay. Clearly, the event ω_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 $\mathcal{A}$. $\qquad \square$

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:

$\qquad \square$
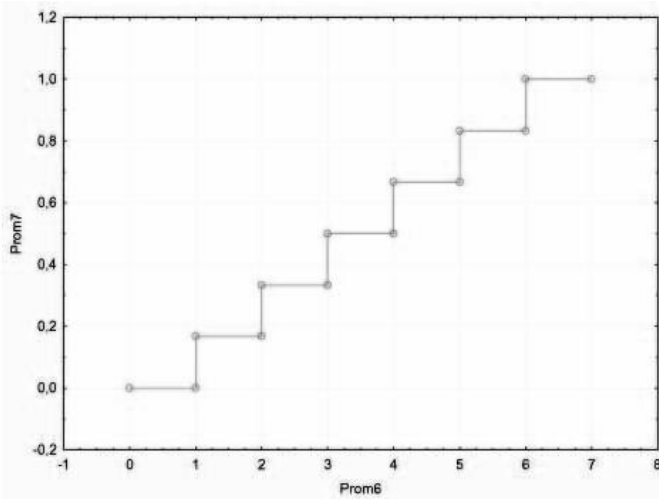
10.E.4. An archer keeps shooting at a target until he hits. He has 4 arrows at his disposal. In each attempt, the probability that he hits the target is 0.6. Let X be the random variable which gives the number of unused arrows. Find the probability mass function and the distribution function of X and 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 end, no matter whether he hits the last time or not. Thus, $f_X(0) = P(X = 0) = 0.4^3$.

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 ± 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 6.3.15 (page 431). 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.

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:

Bodový graf z Prom7 proti Prom6
Tabulka1 10v*30c

Tabulka1 10v*30c

$\square$
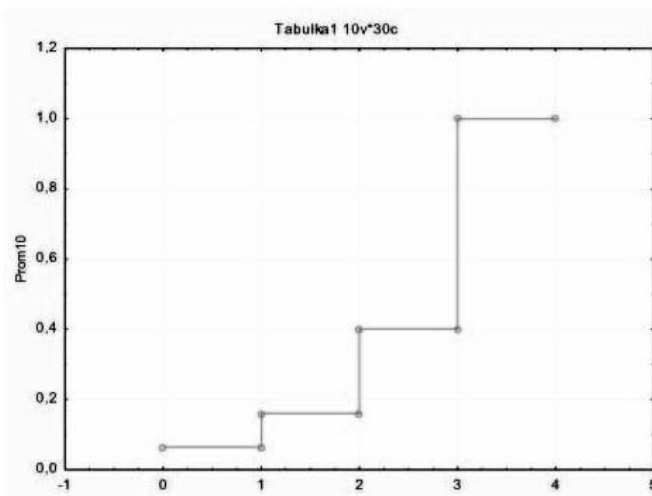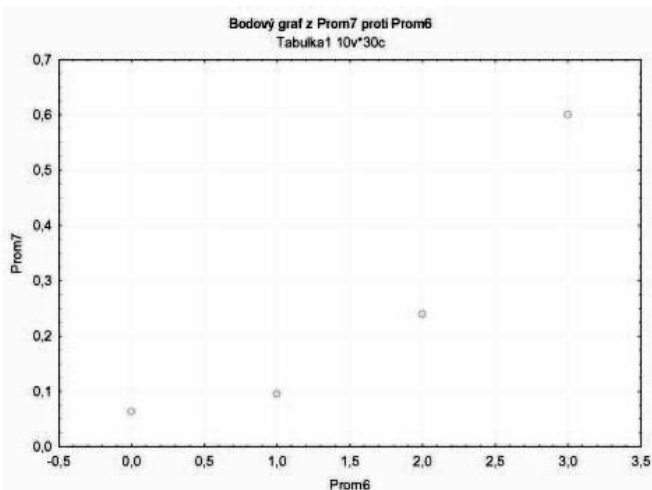
10.E.5. The distribution function of a random variable X is

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 3 \\ \frac{1}{3}x - 1 & \text{for } 3 < x \leq 6 \\ 1 & \text{for } 6 < x. \end{cases}$$

i) Justify that F_X is indeed a distribution function.

DEGENERATE DISTRIBUTION

The distribution which corresponds to a constant random variable $X = \mu$ is called the *degenerate distribution* $\mathrm{Dg}(\mu)$.

Its distribution function F_X and probability mass function f_X are given by

$$F_X(t) = \begin{cases} 0 & t \leq \mu \\ 1 & t > \mu \end{cases} \qquad f_X(t) = \begin{cases} 1 & t = \mu \\ 0 & \text{otherwise.} \end{cases}$$

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* $\mathrm{A}(p)$.

Its distribution function F_X and probability mass function f_X are given by

$$F_X(t) = \begin{cases} 0 & t \leq 0 \\ q & 0 < t \leq 1 \\ 1 & t > 1 \end{cases} \qquad f_X(t) = \begin{cases} p & t = 1 \\ q & t = 0 \\ 0 & t \notin \{0,1\}. \end{cases}$$

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 $t = 0, \ldots, n$, 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 the the binomial distribution of X:

BINOMIAL DISTRIBUTION

The binomial distribution $\mathrm{Bi}(n, p)$ has probability mass function

$$f_X(t) = \begin{cases} \binom{n}{t} p^t (1-p)^{n-t} & t \in \{0, 1, \ldots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

The illustration shows the probability mass functions for $\mathrm{Bi}(50, 0.2)$, and $\mathrm{Bi}(50, 0.9)$. The distribution of the probability corresponds to the intuition that most outcomes occur near the value np:

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 $\lim\limits_{x \to -\infty} F(x) = 0$ and $\lim\limits_{x \to \infty} F(x) = 1$, 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 $P(2 < X < 4) = F_X(4) - F_X(2) = \frac{4}{3} - 1 = \frac{1}{3}$. $\qquad \square$

10.E.6. Consider a random variable X and a function $f : \mathbb{R} \to \mathbb{R}$ 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

$$1 = \int_{-\infty}^{\infty} \frac{a}{1+x^2} dx = a[\operatorname{arctg} x]_{-\infty}^{\infty} = a\pi.$$

Hence $a = \frac{1}{\pi}$.

b) By 10.2.14, the distribution function is given by the following integral:

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\pi} \int_{-\infty}^{x} \frac{dt}{1+t^2} = \frac{1}{\pi} \operatorname{arctg} x + \frac{1}{2}.$$

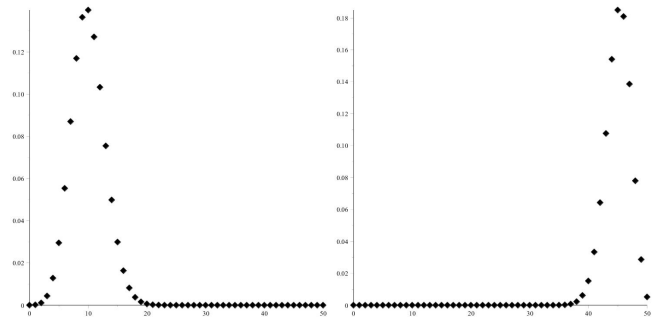c) By b) and the definition of the distribution function, we have

$$P(-1 < X < 1) = F_X(1) - F_X(-1) = \frac{1}{\pi} \cdot \frac{\pi}{4} - \frac{1}{\pi} \cdot \left(-\frac{\pi}{4}\right) = \frac{1}{2}. \qquad \square$$

10.E.7. The joint probability mass function of a discrete random vector is given by the following table:

X \ Y	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;

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 $\operatorname{Bi}(r, 1/n)$.

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.

726

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 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 $X \le a$ and $Y \le b$. 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:

$F_{(X,Y)}$	[2,5)	[5,6)	[6,∞)
[1,2)	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{7}{20}$
[2,3)	$\frac{3}{10}$	$\frac{9}{20}$	$\frac{1}{2}$
[3,∞)	$\frac{3}{5}$	$\frac{4}{5}$	1

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:

$$f_{(X,Y)}(x,y) = \begin{cases} \frac{1}{6}(4x - y) & \text{for } 1 \le x \le 2, 2 \le y \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

Solution. By definition, we have

$$P(2X > Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{2x} f_{(X,Y)}(x,y) dy dx =$$

$$= \int_{1}^{2} \int_{2}^{2x} \frac{1}{6}(4x - y) dy dx =$$

$$= \int_{1}^{2} \left[\frac{2}{3}xy - \frac{1}{12}y^2 \right]_{2}^{2x} dx =$$

$$= \int_{2}^{4} \left(x^2 - \frac{4}{3}x + \frac{1}{3} \right) dx =$$

$$= \left[\frac{1}{3}x^3 - \frac{2}{3}x^2 + \frac{1}{3}x \right]_{1}^{2} = \frac{2}{3}.$$

□

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 λ of elements, on average.

We are interested in the asymptotic behaviour of the variables X_n as $n \to \infty$. Letting $\lim_{n\to\infty} r_n/n = \lambda$, 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:

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \binom{r_n}{k} \frac{(n-1)^{r_n - k}}{n^{r_n}}$$

$$= \lim_{n\to\infty} \frac{r_n(r_n - 1)\dots(r_n - k + 1)}{(n-1)^k} \frac{1}{k!} \left(1 - \frac{1}{n}\right)^{r_n}$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1 + \frac{-\frac{r_n}{n}}{r_n}\right)^{r_n} = \frac{\lambda^k}{k!} \lim_{m\to\infty} \left(1 + \frac{-\lambda}{m}\right)^{m}$$

$$= \frac{\lambda^k}{k!} e^{-\lambda},$$

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* $\mathrm{Po}(\lambda)$ 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,

$$\sum_{k=0}^{\infty} f_X(k) = \sum_{k} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k} \frac{\lambda^k}{k!} = e^{-\lambda + \lambda} = 1.$$

As seen above, this discrete distribution $\mathrm{Po}(\lambda)$ with an arbitrary $\lambda > 0$ (distributed into infinitely many points) is a good approximation of the binomial distributions $\mathrm{Bi}(n, \lambda/n)$, 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 λ 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.

□

10.E.9. Find the marginal distribution function and the joint and marginal density of the random vector (X, Y) provided

$$F_{(X,Y)}(x, y) = \begin{cases} 0 & \text{for } x < 0, y < 0 \\ \frac{1}{4}x^2y^2 & \text{for } 0 \le x \le 1, 0 \le y \le 2 \\ 1 & \text{for } x > 1, y > 2 \end{cases}$$

Solution. The density of the random vector (X, Y) is obtained by differentiation with respect to x a y. Thus, for $0 \le x \le 1, 0 \le y \le 2$, we have $f_{(X,Y)}(x, y) = xy$, and elsewhere the density is zero. The marginal density of the random variable X is then

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)dy = \int_0^2 xy\,dy = [\frac{1}{2}xy^2]_0^2 = 2x.$$

Similarly, for Y, we get $f_Y(y) = \frac{1}{2}y$. The marginal distribution functions are

$$F_X(x) = \int_{-\infty}^x f_X(t)dt = \int_0^x 2t\,dt = x^2$$

and

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt = \int_0^y \frac{1}{2}t\,dt = \frac{1}{4}y^2. \qquad \square$$

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 $P(X \le 3), P(1 \le Y \le 4)$.

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\y	0	1	2	3	4	5	Σ_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
Σ_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 second example is more challenging. We describe events which occur randomly at time $t \ge 0$. 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 $\lim_{h \to 0_+} \frac{\alpha(h)}{h} = 0$;
- the probability $\beta(h)$ of more than one event occurring in a time period of length h to satisfy $\lim_{h \to 0_+} \frac{\beta(h)}{h} = 0$;
- 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,

$$p_k(t + h) = P(X_t = k, X_{t+h} - X_t = 0)$$
$$+ P(X_t = k - 1, X_{t+h} - X_t = 1)$$
$$+ P(X_t \le k - 2, X_{t+h} = k)$$
$$= p_k(t)P(X_{t+h} - X_t = 0) + p_{k-1}P(X_{t+h} - X_t = 1)$$
$$+ \sum_{i=0}^{k-2} P(X_t = i, X_{t+h} - X_t = k - i)$$
$$= p_k(t)(1 - h\lambda - \alpha(h) - \beta(h)) + p_{k-1}(t)(h\lambda + \alpha(h))$$
$$+ \sum_{i=0}^{k-2} p_i(t)P(X_{t+h} - X_t = k - i).$$

Hence (similar to in 6.1.16, page 389, the symbol $\mathrm{o}(h)$ is written for expressions which, when divided by h, approach zero as $h \to 0_+$)

$$\frac{p_0(t + h) - p_0(t)}{h} = -\lambda p_0(t) + \frac{1}{h}\mathrm{o}(h)$$
$$\frac{p_k(t + h) - p_k(t)}{h} = -\lambda p_k(t) + \lambda p_{k-1}(t) + \frac{1}{h}\mathrm{o}(h).$$

Letting $h \to 0_+$, an (infinite!) system of ordinary differential equations is obtained:

$$p_0'(t) = -\lambda p_0(t), \quad p_0(0) = 1$$
$$p_k'(t) = -\lambda p_k(t) + \lambda p_{k-1}(t), \quad p_k(0) = 0$$

for all $t > 0$ and $k \in \mathbb{N}$, with an initial condition.

The first equation has a unique solution

$$p_0(t) = \mathrm{e}^{-\lambda t},$$

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

$$P(1 \leq Y \leq 4) = F_Y(4) - F_Y(1) =$$
$$= \frac{1}{3003}(630 + 1260 + 840 + 180) = 0.969.$$

$\square$

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

$$1 = \int_0^1 \int_0^1 \int_0^1 c(x+y+z)dzdydx = c\int_0^1 \int_0^1 (x+y+\tfrac{1}{2})dydx =$$
$$= c\int_0^1 (x+1)dx = \tfrac{3}{2}c.$$

Hence, $c = \frac{2}{3}$. By definition, the distribution function is equal to

$$F_X(x,y,z) = \tfrac{2}{3}\int_0^x \int_0^y \int_0^z (r+s+t)dtdsdr =$$
$$= \tfrac{2}{3}\int_0^x \int_0^y (rz + sz + \tfrac{1}{2}z^2)dsdr = \tfrac{2}{3}\int_0^x (rzy + \tfrac{1}{2}y^2z + \tfrac{1}{2}z^2y)dr =$$
$$= \tfrac{2}{3}(\tfrac{1}{2}x^2zy + \tfrac{1}{2}y^2zx + \tfrac{1}{2}z^2yx) = \tfrac{1}{3}(x^2zy + y^2zx + z^2yx),$$

so the wanted probability is

$$P(0 \leq X \leq \tfrac{1}{2}, 0 \leq Y \leq \tfrac{1}{2}, 0 \leq Z \leq \tfrac{1}{2}) = F(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) = \tfrac{1}{16}. \qquad \square$$

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

$$\int_{-\infty}^{\infty} f(x) = 1$$

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,

$$\int_{-\infty}^{\infty} f(x) = \int_1^2 a\ln(x) = a[x(\ln(x)-1)]_1^2 = a(2\ln(2)-1),$$

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

$$p_k(t) = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t},\ t > 0,\ k \in \mathbb{N}.$$

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 $\mathrm{Po}(\lambda t)$.

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 $s \to 0_+$ (and applying l'Hospital's rule),

$$\big(\ln(p)\big)'(t) = \lim_{s \to 0_+} \frac{\ln p(t+s) - \ln p(t)}{s}$$
$$= \lim_{s \to 0_+} \frac{(\ln p(s))'}{1} = \frac{p'(0)}{p(0)} = p'(0).$$

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

$$p(t) = e^{-\lambda t}.$$

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 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.

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* $\mathrm{ex}(\lambda)$.

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:

$$\int_0^\infty cx^{a-1}e^{-bx}\,dx = \int_0^\infty c\left(\frac{t}{b}\right)^{a-1}e^{-t}\frac{1}{b}\,dt = \frac{c}{b^a}\Gamma(a).$$

Γ is the famous transcendental function providing the analytic extension of the factorial function, discussed in 6.2.17 on the page 410.

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 $\mathrm{Bi}(n, p)$ for large values of n.

Recall the smooth function $y = e^{-x^2/2}$, mentioned in subsection 6.1.5 (page 376) as an appropriate tool for the construction of functions which are smooth but not analytic. The

Solution. First, we find the distribution function F (for $0 < d < \frac{4}{3}\pi r^3$)

$$F(d) = P\left[\frac{4}{3}\pi X^3 \le d\right] = P\left[X \le \sqrt[3]{\frac{3d}{4\pi}}\right] = \frac{\sqrt[3]{\frac{3d}{4\pi}}}{r}.$$

Altogether,

$$F(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ \sqrt[3]{\frac{3}{4\pi r^3}}\, x^{\frac{1}{3}} & \text{for} \quad 0 < x < \frac{4}{3}\pi r^3 \\ 1 & \text{for} \quad x \ge \frac{4}{3}\pi r^3 \end{cases}$$

Differentiating this, we obtain the density:

$$f(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ \sqrt[3]{\frac{1}{36\pi r^3}}\, x^{-\frac{2}{3}} & \text{for} \quad 0 < x < \frac{4}{3}\pi r^3 \\ 0 & \text{for} \quad x \ge \frac{4}{3}\pi r^3 \end{cases}$$

$\square$

10.E.15. Find the value(s) of the parameter $a \in \mathbb{R}$ so that the function

$$f(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ ax^2 & \text{for} \quad 0 < x < 3 \\ 0 & \text{for} \quad x \ge 3 \end{cases}$$

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 $t \in (0, 27)$ 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

$$f(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ ax & \text{for} \quad 0 < x < 3 \\ 0 & \text{for} \quad x \ge 3 \end{cases}$$

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.

illustration compares this curve (in the right hand part) to the values of $\mathrm{Bi}(40, 0.5)$.
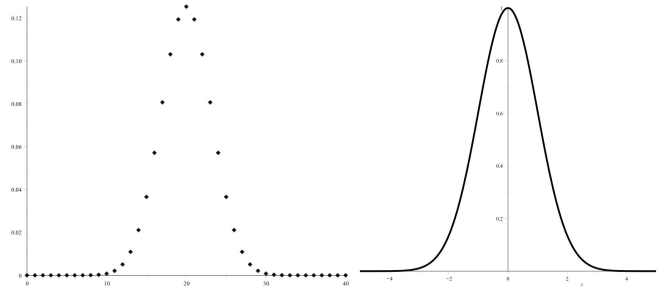
This suggests looking for a convenient continuous distribution whose density would be given by a suitably adjusted variation of this function.

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

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right)$$

$$= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\, dxdy$$

$$= \int_0^\infty \int_0^{2\pi} e^{-(r^2)/2}\, rdrd\theta$$

$$= 2\pi$$

(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* $\mathrm{N}(0, 1)$. 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

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. $\qquad\square$

10.E.17. Find the value(s) of the parameter $a \in \mathbb{R}$ so that the function

$$f(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ ax^2 & \text{for} \quad 0 < x < 2 \\ 0 & \text{for} \quad x \geq 2 \end{cases}$$

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. $\qquad\bigcirc$

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 $\frac{l + \sqrt{l^2 - 4d}}{2}$. The inequality is satisfied by exactly those values of x which lie outside this interval. Therefore,

$$P[X(l - X) \leq d] = P[X \in \langle 0, l \rangle \setminus (\frac{l - \sqrt{l^2 - 4d}}{2}, \frac{l + \sqrt{l^2 - 4d}}{2})]$$
$$= \frac{l - \sqrt{l^2 - 4d}}{l} = 1 - \frac{\sqrt{l^2 - 4d}}{l}.$$

shift μ of the variable z, while scaling the difference $x - \mu$ by coefficient $\sigma > 0$ does the rest. Thus, there are two real parameters μ 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 $\mathrm{N}(\mu, \sigma^2)$.

We return to the asymptotic closeness of the normal and binomial distributions for $n \to \infty$ after creating suitable tools. The following illustration reveals, how well this works. The discrete values correspond to $\mathrm{Bi}(40, 0.5)$, while the curve depicts the density of $\mathrm{N}(20, 10)$.
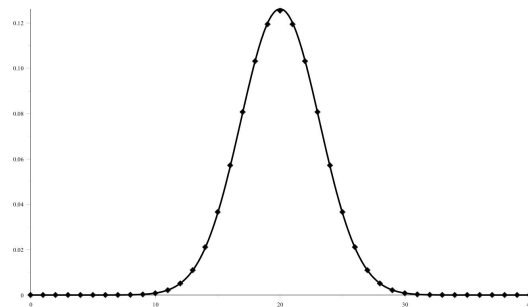
10.2.22. Distributions of random vectors. As for the scalar random variables, one defines the distribution functions and the density or the probability mass function for continuous and discrete random vectors. There are joint probability mass functions and densities.

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 $a, b \in \mathbb{R}$,

$$F(a,b) = P(X < a, Y < b) =$$
$$= \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)dxdy,$$

and the function $f(x,y)$ is called the *(joint) density* of the random vector (X, Y).

Altogether,

$$F(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ 1 - \frac{\sqrt{l^2 - 4x}}{l} & \text{for} \quad 0 \le x \le \frac{l^2}{4} \\ 1 & \text{for} \quad x > \frac{l^2}{4} \end{cases}$$

The density is obtained by differentiation:

$$f(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ \frac{2}{l\sqrt{l^2 - 4x}} & \text{for} \quad 0 \le x \le \frac{l^2}{4} \\ 0 & \text{for} \quad x > \frac{l^2}{4} \end{cases}$$

□

10.E.19. Nezávislé náhodné veličiny X a Y mají následující hustoty pravděpodobnosti:

$$f_X(t) = \begin{cases} 0 & \text{for } t \le 0, \\ 1 & \text{for } 0 < t < 1, \\ 0 & \text{for } 1 \le t, \end{cases} \quad f_Y(t) = \begin{cases} 0 & \text{for } t \le 0, \\ 2t & \text{for } 0 < t < 1, \\ 0 & \text{for } 1 \le t. \end{cases}$$

Určete distribuční funkci náhodné veličiny udávající obsah obdélníka o stranách X a Y.

Solution.

$$F_Y(t) = \begin{cases} 0 & \text{for} \quad t \le 0 \\ 2t - t^2 & \text{for} \quad 0 < t < 1 \\ 1 & \text{for} \quad 1 \le t \end{cases}$$

□

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:

$$f(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{for} \quad x \ge 1. \end{cases}$$

Find the probability that Y is less than X^2.

Solution. Since X and Y are independent random variables, the joint density $f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}^2$ of the variable (X, Y) is given by the densities f_X and f_Y of the individual random variables. Thus, we have

$$f_{(X,Y)}(u, v) =$$
$$= \begin{cases} f_X(u) \cdot f_Y(v) = \frac{1}{2} \cdot 2v = v & \text{for } (u, v) \in (0, 2) \times (0, 1), \\ 0 & \text{otherwise.} \end{cases}$$

Then, the wanted probability P is the integral of the density $f_{(X,Y)}$ over the part O of the plane where $Y < X^2$:

$$P = \iint\limits_O f_{(X,Y)} \, \mathrm{d}x \, \mathrm{d}y = 1 - \iint\limits_{\mathbb{R}^2 \setminus O} f_{(X,Y)} \, \mathrm{d}x \, \mathrm{d}y =$$

$$= 1 - \int_0^1 \int_{x^2}^1 y \, \mathrm{d}y \, \mathrm{d}x = \frac{3}{5}.$$

□

For a general continuous random vector $X = (X_1, \ldots, X_n)$, define

$$F(a_1, \ldots, a_n) = P(X_1 < a_1, \ldots, X_n < a_n) =$$
$$= \int_{-\infty}^{a_n} \cdots \int_{-\infty}^{a_1} f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n,$$

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 $X_1, \cdots, X_n$ are *(stochastically) independent* if and only if for any $a_i \in \mathbb{R}$, the events $X_1 < a_1$, ..., $X_n < a_n$ are independent. In view of the definition of the distribution function F of the random vector $(X_1, \ldots, X_n)$, this is equivalent to

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n),$$

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 $F(b, c_2, \ldots, c_n) - F(a, c_2, \ldots, c_n) = (F_{X_1}(b) - F_{X_1}(a))F_{X_2}(c_2) \ldots F_{X_n}(c_n)$, and so on. The densities and probability mass functions behave well too:

Proposition. *For any random vector $(X_1, \ldots, X_n)$, the following two conditions are equivalent:*
- *The random variables $X_1 \ldots, X_n$ are stochastically independent.*
- *The joint distribution function F of the of random vector $(X_1, \ldots, X_n)$ is the product of the marginal distribution functions F_{X_i} of the individual components.*

10.E.21. Let X, Y be independent random variables, where X has density function

$$f_1(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ 2x & \text{for} \quad 0 < x < 1 \\ 0 & \text{for} \quad x \geq 1, \end{cases}$$

and Y has density function

$$f_2(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ \frac{x}{2} & \text{for} \quad 0 < x < 2 \\ 0 & \text{for} \quad x \geq 2. \end{cases}$$

Find the probability that Y is greater than X^2. ◯

Solution. $f_{(X,Y)}(u, v) = uv$, for $(u, v) \in (0, 1) \times (0, 2)$, $f_{(X,Y)}(u, v) = 0$ otherwise. Then, the wanted probability is

$$P = \int_0^1 \int_{x^2}^2 xy \, \mathrm{d}y \, \mathrm{d}x = \frac{11}{12}. \qquad \square$$

10.E.22. Let X, Y be independent random variables, where X has density function

$$f(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ \frac{2x}{9} & \text{for} \quad 0 < x < 3 \\ 0 & \text{for} \quad x \geq 1, \end{cases}$$

and Y has density function

$$f(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ \frac{x}{2} & \text{for} \quad 0 < x < 2 \\ 0 & \text{for} \quad x \geq 2. \end{cases}$$

Find the probability that Y is greater than X^3.

◯

Solution.

$$P = \int_0^{\sqrt[3]{2}} \int_{x^3}^2 xy \, \mathrm{d}y \, \mathrm{d}x = \frac{\sqrt[3]{4}}{12}. \qquad \square$$

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 $Y_1, \ldots, Y_n \sim A(p)$ are independent random variables saying whether the k–th experiment was successful. Clearly, the Bernoulli distribution has expected value $\mathrm{E}\, Y_i = p$, hence by theorem 10.2.29, we have $\mathrm{E}\, X = \sum_{k=1}^n \mathrm{E}\, Y_k = np$. Similarly, we compute $\mathrm{E}(Y_k^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$, so $\text{var}\, Y_k = \mathrm{E}(Y_k^2) - (\mathrm{E}\, Y_k)^2 = p - p^2$. By theorem 10.2.33, we have $\text{var}\, X = \sum_{k=1}^n \text{var}\, Y_k = np(1 - p)$. $\square$

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 $(X_1, \ldots, X_n)$ 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 $(X_1, \ldots, X_n)$ 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 $f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_j) = f_X(x_i)f_Y(y_j)$. The joint distribution function is

$$F_{X,Y}(x, y) = \sum_{x_i < x} \sum_{y_j < y} f_X(x_i)f_Y(y_j)$$

$$= \left(\sum_{x_i < x} f_X(x_i) \right) \left(\sum_{y_j < y} f_Y(y_j) \right),$$

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:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

$$= \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y)$$

$$= f_X(x)f_Y(y),$$

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. $\square$

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

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 $X \sim \mathrm{Bi}(5, \frac{3}{5})$. By F, the expected value and variance of $\mathrm{Bi}(n, p)$ are equal to np and $np(1-p)$, respectively, which gives $\mathrm{E}\, X = 3$ and $\mathrm{var}\, X = \frac{6}{5}$ for our case. $\qquad\square$

10.F.2. Consider the discrete random variable X which takes on the values $k = 0, 1, 2, 3, \ldots$, each with probability $P(X = k) = p(1 - p)^k$ (geometric distribution). Find $\mathrm{E}\, X$ (the expected number of failures before the first success) and $\mathrm{var}\, X$.

Solution. Using the definition of the expected value and the formula for summing the derivative of a geometric series, we calculate

$$\mathrm{E}\, X = \sum_{k=0}^{\infty} kp(1-p)^k = p(1-p) \sum_{k=0}^{\infty} k(1-p)^{k-1} =$$
$$= p(1-p)\frac{1}{p^2} = \frac{1-p}{p}.$$

Similarly, using the formula for summing the second derivative of a geometric series, we compute

$$E(X^2) \quad = \quad \sum_{k=0}^{\infty} k^2 p(1-p)^k = \frac{(1-p)(2-p)}{p^2},$$

hence the variance is $\mathrm{var}\, X = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2 = \frac{1-p}{p^2}$. $\quad\square$

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

$$\mathrm{E}\, X = \int_1^{\infty} \frac{3}{x^3} dx = \left[-\frac{3}{2x^2}\right]_1^{\infty} = \frac{3}{2}$$

and the expected value of X^2 is

$$\mathrm{E}(X^2) = \int_1^{\infty} \frac{3}{x^2} dx = \left[-\frac{3}{x}\right]_1^{\infty} = 3.$$

Therefore, $\mathrm{var}\, X = 3 - (\frac{3}{2})^2 = \frac{3}{4}$. $\qquad\square$

$(\pm 1, \pm 1)$ 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, Φ),

$$P(R < r_0, \Phi < \varphi_0) = \int_0^{r_0} \int_0^{\varphi_0} \frac{1}{\pi} r \, d\varphi dr = \frac{1}{2\pi} \varphi_0 r_0^2.$$

The joint density of the vector (R, Φ) is thus $f(r, \varphi) = \frac{r}{\pi}$ for $0 < r \le 1, 0 < \varphi \le 2\pi$, and it is zero otherwise. The marginal densities are

$$f_R(r) = \int_0^{2\pi} \frac{r}{\pi} d\varphi = 2r, \quad \text{if } 0 < r \le 1,$$

$$f_\Phi(\varphi) = \int_0^1 \frac{r}{\pi} dr = \frac{1}{2\pi}, \quad \text{if } 0 < \varphi \le 2\pi,$$

and zero otherwise. Therefore, the random variables R and Φ 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 $(X_1, \ldots, X_n)$ 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 ψ is to guarantee that $Y = \psi \circ X$ is again a random variable according to the definition. This means the preimage ψ^{-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 ψ and open sets U, the continuous function ψ always satisfies the condition.

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

$$\mathrm{E}\, X = \int_0^{\frac{\pi}{2}} x \cos x\, dx = [x \sin x + \cos x]_0^{\frac{\pi}{2}} = \frac{\pi}{2} - 1.$$

Using double integration by parts, we obtain

$$\mathrm{E}(X^2) = \int_0^{\frac{\pi}{2}} x^2 \cos x\, dx =$$

$$= [x^2 \sin x + 2x \cos x - 2 \sin x]_0^{\frac{\pi}{2}} = \left(\frac{\pi}{2}\right)^2 - 2,$$

so the variance is equal to $\mathrm{var}\, X = (\frac{\pi}{2})^2 - 2 - (\frac{\pi}{2}-1)^2 = \pi - 3$. By definition, the distribution function is equal to $F_X(x) = \int_0^x \cos t\, dt = \sin x$, and the median is $F^{-1}(0.5) = \frac{\pi}{6}$. □

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

$$\mathrm{E}\, X = \int_0^\infty x \lambda e^{-\lambda x}\, dx = \left[-x e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda},$$

$$\mathrm{E}(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx =$$

$$= \left[-x^2 e^{-\lambda x} - 2x \frac{1}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \frac{2}{\lambda^2},$$

hence $\mathrm{var}\, X = \mathrm{E}(X^2) - (\mathrm{E}\, X)^2 = \frac{1}{\lambda^2}$. 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 $F^{-1}(0.5) = -\frac{1}{\lambda} \ln(\frac{1}{2}) = \frac{\ln 2}{\lambda}$. □

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 $\mathrm{cov}(X_1, X_2)$.

Solution. If π is to be a probability mass function, then the sum of its values over the entire domain must be equal to 1, i.e.,

$$\sum_{i,j} \pi(i, j) = c + 3.2c + 3c = 10c = 1,$$

In the sequel, we restrict ourselves to this case.

Functions of random variables and vectors

For a continuous function $\psi : \mathbb{R} \to \mathbb{R}$ 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 $(X_1, \ldots, X_n)$ and a continuous function $\psi : \mathbb{R}^n \to \mathbb{R}$, we talk about a function $Y = \psi(X_1, \ldots, X_n)$ 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

$$A_u = \{x;\ g(x) < u\} \qquad B_v = \{y;\ h(y) < v\}.$$

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

$$(1) \qquad f_{\psi(X)}(y) = P(\psi(X) = y) = \sum_{\psi(x_i) = y} f(x_i).$$

Thus, in the case of the affine dependency $Y = a + bX$, the probability mass function is non-zero exactly at the points $y_i = a x_i + b$.

As an example of a function of a random vector $X = (X_1, \ldots, X_n)$, consider the sum of n independent random variables with the Bernoulli distribution $X_i \sim \mathrm{A}(p)$. Of course, this leads just to the binomial distribution $\mathrm{Bi}(n, p)$. The above formula for $f_{\psi(X)}(y)$ reveals the already known probability function for $Y = X_1 + \cdots + X_n$. Only $y \in \{0, \ldots, n\}$ 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

so $c = \frac{1}{10}$. The probability mass function π_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 the probability mass function π_2 of X_2, we get $\pi_2(-1) = c$, $\pi_2(0) = 5c, \pi_2(1) = 4c$ and zero elsewhere. Hence, $\mathrm{E}\, X_1 = \sum_i i\pi_1(i) = 14c = 1.4$ and $\mathrm{E}\, X_2 = \sum_j j\pi_2(j) = = 3c = 0.3$. By the definition of the covariance, we have

$$\mathrm{cov}(X_1, X_2) = \sum_{i,j}(i - 1, 4)(j - 0, 3)\pi(i, j) = 0.18.$$

$\square$

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 $[0, 1]$:

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1},$$

where α, β are fixed parameters, chosen suitably for description of the given random variable, and $\mathrm{B}(\alpha, \beta)$ 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

$$\mathrm{E}\, X = \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 x^{\alpha}(1 - x)^{\beta-1} dx.$$

Integrating by parts, we get

$$\mathrm{E}\, X = -\frac{1}{\mathrm{B}(\alpha, \beta)\beta}[x^{\alpha}(1-x)^{\beta}]_0^1 + \frac{\alpha}{\mathrm{B}(\alpha, \beta)\beta} \int_0^1 x^{\alpha-1}(1-x)^{\beta} dx.$$

Clearly, the first term is equal to zero. Refining the second one, we obtain

$$\mathrm{E}\, X = \frac{\alpha}{\mathrm{B}(\alpha, \beta)\beta} \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx -$$

$$- \frac{\alpha}{\mathrm{B}(\alpha, \beta)\beta} \int_0^1 x^{\alpha}(1 - x)^{\beta-1} dx.$$

Now, the integral in the first term is, thanks to the normalization, equal to $\mathrm{B}(\alpha, \beta)$, and the second integral is the expected

$Z \sim \mathrm{N}(0, 1)$ in 10.2.21. This is verified easily.

$$F_Y(y) = P(Y < y) = P(\mu + \sigma Z < y)$$

$$= \Phi\left(\frac{1}{\sigma}(y - \mu)\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-\mu}{\sigma}} e^{-z^2/2}\, dz$$

$$= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx,$$

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 ψ has got non-zero derivative (thus ψ is invertible).

$$(2) \qquad f_Y(y) = \int_{-\infty}^{y} |\psi'(\psi^{-1}(y)|^{-1} f_X(\psi^{-1}y)\, dy.$$

Check the formula yourself! (Start with the case when the derivative ψ' 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):

$$F_V(u) = \int_{x+y<u} f_X(x)f_Y(y)\, dxdy =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{u-x} f_X(x)f_Y(y)\, dydx$$

$$= \int_{-\infty}^{u} \left(\int_{-\infty}^{\infty} f_X(x)f_Y(v - x)\, dx \right) dv,$$

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

$$(3) \qquad f_V = f_X * f_Y,$$

already met in subsection 7.2.2 (page 472). 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 blury 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 deviationare looked for. Now we introduce similar characteristics for random variables and

737

value, too. Thus, the above equation can be written as

$$\mathrm{E}\,X = \frac{\alpha}{\beta} - \frac{\alpha}{\beta}\,\mathrm{E}\,X.$$

Hence it immediately follows that $\mathrm{E}\,X = \frac{\alpha}{\alpha+\beta}$.

c) In order to compute the variance, we must calculate

$$\mathrm{E}\,X^2 = \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}dx.$$

This integral can be computed similarly as in b):

$$\mathrm{E}\,X^2 = \frac{\alpha+1}{\mathrm{B}(\alpha,\beta)\beta}\int_0^1 x^{\alpha}(1-x)^{\beta}dx = \frac{\alpha+1}{\beta}\,\mathrm{E}\,X - \frac{\alpha+1}{\beta}\,\mathrm{E}\,X^2.$$

Hence, $\mathrm{E}\,X^2 = \frac{(\alpha+1)\,\mathrm{E}\,X}{\alpha+\beta+1}$. Substituting the expected value, we obtain

$$\mathrm{var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \qquad \square$$

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:

X \ Y	0	1	2
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:

$$E(X) = E(Y) = 0\cdot\frac{1}{4} + 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} = 1$$

$$E(X^2) = E(Y^2) = 0\cdot\frac{1}{4} + 1\cdot\frac{1}{2} + 4\cdot\frac{1}{4} = \frac{3}{2}$$

$$E(XY) = 0\cdot\frac{3}{8} + 1\cdot\frac{1}{4} + 2\cdot\frac{1}{4} + 4\cdot\frac{1}{8} = \frac{5}{4}$$

Thus, we have

$$\sigma^2(X) = \varsigma^2(Y) = E(X^2) - [E(X)]^2 = \frac{1}{2}$$

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{4}$$

random vectors. The first one is an analogy of the arithmetic mean.

EXPECTED VALUE

For any random variable X, define its *expected value* $\mathrm{E}\,X$ by

$$\mathrm{E}\,X = \begin{cases} \sum_i x_i f_X(x_i) & \text{for a discrete variable} \\ \int_{-\infty}^{\infty} x f_X(x)dx & \text{for a continuous variable,} \end{cases}$$

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 ψ for which Y is again a random variable.

In the discrete case, compute

$$\mathrm{E}\,Y = \sum_j y_j P(Y = y_j)$$
$$= \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_j)$$
$$= \sum_i \psi(x_i) P(X = x_i),$$

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 :

$$\mathrm{E}\,\psi(X) = \int_{-\infty}^{\infty} \psi(x) f_X(x)dx,$$

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 ψ is a differentiable monotone function with non-zero derivative, it is an easy exercise to verify that the definition of $\mathrm{E}\,\psi(X)$ coincides with $\mathrm{E}\,Y$. 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

Altogether,

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{1}{2}$$

$\square$

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

Find the marginal distributions of both variables, their expected values, variances, and correlation coefficient. $\bigcirc$

10.F.10. Find the expected value and variance of the random variable X^2 provided X has uniform distribution on the interval $\langle -1, 1 \rangle$. $\bigcirc$

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. $\bigcirc$

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. $\bigcirc$

G. Transformations of random variables

Consider a continuous function of a random variable $Y = \psi(X)$. If the transformation ψ is increasing, then the resulting distribution function satisfies

$$F_Y(y) = P(Y \le y) = P(\psi(X) \le y) =$$
$$= P(X \le \psi^{-1}(y)) = F_X(\psi^{-1}(y)),$$

where F_X is the distribution function of X (analogously for decreasing ψ). 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

$$\mathrm{E}\,Y = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx,$$

ans similarly for the variance of Y.

is the random variable corresponding to the won amount, it seems that the correct answer is "anything below the expected value $\mathrm{E}\,X$".

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.[3]

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 $\mathrm{A}(p)$, it is immediate that

$$\mathrm{E}\,X = (1 - p) \cdot 0 + p \cdot 1 = p.$$

Similarly, compute the expected value np of the binomial distribution $\mathrm{Bi}(n, p)$. This requires more thought. The result is a direct corollary of the following general theorem since $\mathrm{Bi}(n, p)$ is the sum of n random variables with the Bernoulli distributions $\mathrm{A}(p)$.

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 $\mathrm{E}\,X$ and $\mathrm{E}\,Y$ exist.

It follows directly from the definition that the constant random variable a has expected value a. Further,

$$\mathrm{E}(bX) = b\,\mathrm{E}\,X,$$

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,

$$\mathrm{E}(XY) = \sum_i \sum_j x_i y_j f_X(x_i) f_Y(y_j)$$

$$= \left(\sum_i x_i f_X(x_i) \right) \left(\sum_j y_j f_Y(y_j) \right) = \mathrm{E}\,X\,\mathrm{E}\,Y.$$

Similarly, verify the equality $\mathrm{E}(XY) = \mathrm{E}\,X\,\mathrm{E}\,Y$ for independent continuous random variables.

---

[3]Going back to Bernoulli, 1738, the real value is given by the utility, rather than the price.

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) $f_Y(y) = f(1/y)\frac{1}{y^2}$. $\qquad\square$

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 $Y = \sin X, Z = \operatorname{tg} X$.

Solution. Since the length of the interval where X is nonzero is π, 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

$$f_Z(y) = f_X(\arctan(z))\arctan'(y) = \frac{1}{\pi(1+y^2)}.$$

$\square$

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 $\mathrm{E}\,Y$ and $\operatorname{var} Y$.

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 $\mathrm{E}\,Y = \int_{-\infty}^{\infty} x^2 f_X(x)dx$. Thus,

$$\mathrm{E}\,Y = \int_0^{\frac{\pi}{2}} x^2 \cos x\, dx = \left[x^2 \sin x + 2x\cos x - 2\sin x\right]_0^{\frac{\pi}{2}} =$$

The integral was computed by parts. Applying this method again, we obtain

$$\mathrm{E}(Y^2) = \int_0^{\frac{\pi}{2}} x^4 \cos x\, dx =$$
$$= \left[(x^4 - 12x^2 + 24)\sin x + 4(x^3 - 6x)\cos x\right]_0^{\frac{\pi}{2}}.$$

Now compute $\mathrm{E}(X + Y)$ for arbitrary random variables. For discrete distributions of X and Y,

$$\mathrm{E}(X + Y) = \sum_i \sum_j (x_i + y_j)P(X = x_i, Y = y_j)$$

$$= \sum_i \left(x_i \sum_j P(X = x_i, Y = y_j)\right)$$

$$ + \sum_j \left(y_j \sum_i P(X = x_i, Y = y_j)\right)$$

$$= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j),$$

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.

$$\mathrm{E}(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f_{X,Y}(x, y)\,dxdy$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dydx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dydx$$

$$= \int_{-\infty}^{\infty} x f_X(x)\,dx + \int_{-\infty}^{\infty} y f_Y(y)\,dy = \mathrm{E}\,X + \mathrm{E}\,Y,$$

where absolute convergence of integrals of the expected values $\mathrm{E}\,X$ and $\mathrm{E}\,Y$ is used to interchange the integrals by Fubini's theorem.

Altogether, the expected formula:

$$\mathrm{E}(X + Y) = \mathrm{E}\,X + \mathrm{E}\,Y$$

is obtained, whenever the expected values $\mathrm{E}\,X$ and $\mathrm{E}\,Y$ exist.

Straightforward application of this result leads to the following:

> **AFFINE NATURE OF EXPECTED VALUES**
>
> For any constants $a, b_1, \ldots, b_k$ and random variables $X_1, \ldots, X_k$,
> $$\mathrm{E}(a + b_1 X_1 + \cdots + b_k X_k) = a + b_1\,\mathrm{E}\,X_1 + \cdots + b_k\,\mathrm{E}\,X_k.$$

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 $X = (X_1, \ldots, X_n)$ be a random vector with expected value $\mathrm{E}\,X$, $a \in \mathbb{R}^m$, $B \in \operatorname{Mat}_{mn}(\mathbb{R})$ a martix. Then,*

$$\mathrm{E}(a + B \cdot X) = a + B \cdot \mathrm{E}\,X.$$

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 $\mathrm{E}(a + B \cdot X)$. Thus, it can be assumed that a is a scalar and B is a matrix with a single row.

Hence, $E(Y^2) = (\frac{\pi}{2})^4 - 12(\frac{\pi}{2})^2 + 24$, so $\operatorname{var} Y = \frac{\pi^4}{16} - 3\pi^2 + 24 - \frac{\pi^4 - 16\pi^2 + 64}{16} = 20 - 2\pi^2$. $\square$

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 a $Z = XY$ 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: $EX = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$, $EZ = E(XY) = 0 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = = 0$. 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 $\operatorname{cov}(X, Z) = 0 - \frac{1}{2} \cdot 0 = 0$. 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 $P(Z = 1) \cdot P(X = 0) = \frac{1}{8} \ne 0$. We can see that $P(Z = 1) \cdot P(X = 0) \ne P(Z = 1, X = 0)$, 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 $P(X \ge a) \le \frac{EX}{a}$.

10.H.1. Consider a non-negative random variable X with expected value μ. With no further information about X, bound $P(X > 3\mu)$. Then, compute $P(X > 3\mu)$ if you know that $X \sim \operatorname{Ex}(\frac{1}{\mu})$.

Solution. If the non-negative random variable X does not take zero with probability 1, then its expected value μ is positive. Therefore, the wanted probability can be bounded using Markov's inequality as

$$P(X \ge 3\mu) \le \frac{\mu}{3\mu} = \frac{1}{3}.$$

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. $\square$

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 $(F_X)^{-1} : (0, 1) \to \mathbb{R}$. This means that the value $y = F^{-1}(\alpha)$ is such that $P(X < y) = \alpha$. This corresponds precisely to the quantiles from descriptive statistics using relative frequencies for the probabilities.

<div align="center">QUANTILE FUNCTION</div>

For any random variable X with distribution function $F_X(x)$, define its *quantile function*
$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R}; \ F(x) \ge \alpha\}, \ \alpha \in (0, 1).$$

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 α 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 $\Phi^{-1}(0.975)$, 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.

<div align="center">CRITICAL VALUES</div>

For a random variable X and a real number $0 < \alpha < 1$, define its *critical value* $x(\alpha)$ *at level* α as
$$P(X \ge x(\alpha)) = \alpha.$$
This means that $x(\alpha) = F_X^{-1}(1 - \alpha)$ where F_X^{-1} is the quantile function of the random variable X.

If we know that $X \sim \mathrm{Ex}(\frac{1}{\mu})$, then

$$P(X > 3\mu) = 1 - P(X \le 3\mu) = 1 - F(3\mu),$$

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, $P(X > 3\mu) = \frac{1}{e^3}$. $\qquad\square$

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

$$P(X \le 60) = 1 - P(X \ge 60) \ge 1 - \frac{20}{60} = \frac{2}{3}.$$

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):

$$0.9 \le P(|X - 20| < x) = 1 - P(|X - 20| \ge x) \le 1 - \frac{1}{x^2}.$$

Hence, $x \ge \sqrt{10} \approx 3.2$. Thus, the wanted interval is $(16.8\ \mathrm{km/h}, 23.2\ \mathrm{km/h})$. $\qquad\square$

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, $X \sim \mathrm{Bi}(n, \frac{1}{26})$. We are looking for the value of n for which $P(X \ge 5) = 0.95$, i. e., $F_X(4) = P(X \le 4) = 0.05$. 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 $\mathrm{E}\,X = \frac{n}{26}$, and its variance is $\mathrm{var}\,X = \frac{25n}{26^2}$. Denoting the corresponding

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

$$\mathrm{var}\,X = \mathrm{E}\left((X - \mathrm{E}\,X)^2 \right),$$

provided the right-hand expected value exists. Otherwise, the variance of X does not exist.

The square root $\sqrt{\mathrm{var}\,X}$ 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:

$$\mathrm{var}\,X = \mathrm{E}(X - \mathrm{E}\,X)^2 = \mathrm{E}(X^2 - 2X(\mathrm{E}\,X) + (\mathrm{E}\,X)^2)$$
$$= \mathrm{E}\,X^2 - 2(\mathrm{E}\,X)^2 + (\mathrm{E}\,X)^2$$
$$= \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2.$$

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

$$\mathrm{var}\,Y = \mathrm{E}\big((a + bX) - \mathrm{E}(a + bX)\big)^2 = \mathrm{E}\big(b(X - \mathrm{E}\,X)\big)^2$$
$$= b^2\,\mathrm{var}\,X.$$

Thus are derived the following useful formulae:

PROPERTIES OF VARIANCE

(1)	$\mathrm{var}\,X = E(X^2) - (EX)^2$
(2)	$\mathrm{var}(a + bX) = b^2\,\mathrm{var}\,X$
(3)	$\sqrt{\mathrm{var}(a + bX)} = b\sqrt{\mathrm{var}\,X}$

Given a random variable X with expected value and nonzero variance, define its *standardization* as the random variable

$$Z = \frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}}.$$

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.

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 $\mathrm{N}(0,1)$. 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, $X \sim \mathrm{Bi}(1200, \frac{1}{6})$. By F, we have $\mathrm{E}\,X = 1200 \cdot \frac{1}{6} = 200$ and $\mathrm{var}\,X = 200(1 - \frac{1}{6}) = \frac{500}{3}$. 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,

$$P(150 \leq X \leq 250) = \sum_{k=150}^{250} \binom{1200}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{1200-x}.$$

This expression is hard to evaluate without a computer, so we use Moivre-Laplace theorem. Replacing X with the standardized random variable

$$Z = \frac{\sqrt{3}(X - 200)}{10\sqrt{5}},$$

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 %. $\square$

Theorem. *Consider a random variable X with finite variance, and fix an arbitrary $\varepsilon > 0$. Then,*

$$P(|X - \mathrm{E}\,X| \geq \varepsilon) \leq \frac{\mathrm{var}\,X}{\varepsilon^2}.$$

PROOF. Suppose X is continuous. Set $\mu = \mathrm{E}\,X$ and compute, using the definition:

$$\mathrm{var}\,X = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx$$
$$= \int_{|x-\mu| \geq \varepsilon} (x - \mu)^2 f(x)\,dx$$
$$+ \int_{|x-\mu| < \varepsilon} (x - \mu)^2 f(x)\,dx$$
$$\geq \int_{|x-\mu| \geq \varepsilon} \varepsilon^2 f(x)\,dx = \varepsilon^2\,P(|X - \mu| \geq \varepsilon).$$
$\square$

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 σ, the choice $\varepsilon = k\sigma$ yields the probability

$$P(|X - \mathrm{E}\,X| \geq k\sigma) \leq \frac{1}{k^2}.$$

The Chebyshev's inequality helps understanding asymptotic descriptions of limit processes. For instance, consider the sequence of random variables $X_1, X_2, \ldots$ with probability distributions $X_n \sim \mathrm{Bi}(n, p)$, 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,

$$\mathrm{E}\,Y_n = \frac{np}{n} = p, \quad \mathrm{var}\,Y_n = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

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 $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = 0.$$

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.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 $X \sim \mathrm{Bi}(n, \frac{1}{10})$. By F, we have $\mathrm{E}\,X = 0.1n$ and $\mathrm{var}\,X = 0.09n$. By Chebyshev's inequality 10.2.32, the wanted probability satisfies

$$P(|X - 0.1n| \le 0.02n) = 1 - P(|X - 0.1n| \ge 0.02n) \ge$$
$$\ge 1 - \frac{0.1 \cdot 0.9n}{(0.02n)^2} = 1 - \frac{225}{n}.$$

The inequality $1 - \frac{225}{n} \ge 0.95$ and hence

$$P(|X - 0.1n| \le 0.02n) \ge 0.95$$

holds for $n \ge 4500$. The exact value of the probability is given in terms of the distribution function F_X of the binomial distribution:

$$P(0.08n \le X \le 0.12n) = F_X(0.12n) - F_X(0.08n).$$

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

$$0.95 = P(0.08n \le X \le 0.12n) = P(-\frac{\sqrt{n}}{15} \le Z \le \frac{\sqrt{n}}{15}) \approx$$
$$\approx \Phi(\frac{\sqrt{n}}{15}) - \Phi(-\frac{\sqrt{n}}{15}) =$$
$$= 2\Phi(\frac{\sqrt{n}}{15}) - 1.$$

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. □

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 $X \sim \mathrm{Bi}(500, \frac{4}{5})$. By F, we have $\mathrm{E}\,X = 400$ and $\mathrm{var}\,X = 80$. The standardized random variable is $Z = \frac{X - 400}{\sqrt{80}}$. By the de

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

$$\mathrm{cov}(X, Y) = \mathrm{E}\,((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y))$$

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) $\qquad \mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$
(2) $\qquad \mathrm{cov}(X, Y) = \mathrm{E}(XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y)$
(3) $\qquad \mathrm{cov}(X + Y, Z) = \mathrm{cov}(X, Z) + \mathrm{cov}(Y, Z)$
(4) $\qquad \mathrm{cov}(a + bX, c + dY) = bd\,\mathrm{cov}(X, Y)$
(5) $\qquad \mathrm{var}(X + Y) = \mathrm{var}\,X + \mathrm{var}\,Y + 2\,\mathrm{cov}(X, Y).$

Moreover, if X and Y are independent, then $\mathrm{cov}(X, Y) = 0$, and consequently

(6) $\qquad \mathrm{var}(X + Y) = \mathrm{var}\,X + \mathrm{var}\,Y.$

PROOF. Directly from the definition, the covariance is symmetric in the arguments. The second proposition follows immediately from the properties of the expected value:

$$\mathrm{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)$$
$$= \mathrm{E}(XY) - (\mathrm{E}\,Y)X - (\mathrm{E}\,X)Y + \mathrm{E}\,X\,\mathrm{E}\,Y$$
$$= \mathrm{E}(XY) - \mathrm{E}\,X\,\mathrm{E}\,Y.$$

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:

$$\mathrm{cov}(a + bX, c + dY) =$$
$$= \mathrm{E}\big((a + bX - \mathrm{E}(a + bX))(c + dY - \mathrm{E}(c + dY))\big)$$
$$= \mathrm{E}\big((bX - b\,\mathrm{E}(X))(dY - d\,\mathrm{E}(Y))\big)$$
$$= \mathrm{E}\big(bd(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\big)$$
$$= bd\mathrm{E}\big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\big) = bd\,\mathrm{cov}(X, Y).$$

Moivre-Laplace theorem, we have $F_Z \approx \Phi$, so

$$
\begin{aligned}
P(X \ge 380) &= P(Z \ge \frac{380-400}{\sqrt{80}}) \approx 1 - \Phi(-\frac{\sqrt{20}}{2}) = \\
&= \Phi(\frac{\sqrt{20}}{2}) \approx 0.987.
\end{aligned}
$$

$\square$

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 Φ of the standard normal distribution. Thus, the wanted probability is

$$
\begin{aligned}
P &= 1 - P[759 \le X \le 841] \\
&= 1 - P\left[-2.05 \le \frac{X-800}{20} \le 2.05\right] \\
&\doteq 2\Phi(-2,05) \doteq 0.0404.
\end{aligned}
$$

$\square$

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 $\frac{X-800}{20}$ can be approximated, by the de Moivre-Laplace theorem, with the distribution function Φ of the standard normal distribution. Thus, the wanted probability is

$$
\begin{aligned}
P[1767 \le X \le 1833] &= P\left[-1,1 \le \frac{X-1800}{30} \le 1,1\right] \\
&\doteq \Phi(1,1) - \Phi(-1,1) \doteq 0,7498.
\end{aligned}
$$

$\square$

The other propositions about the variance are quite simple corollaries:

$$
\begin{aligned}
\operatorname{var}(X+Y) &= \operatorname{E}\big((X+Y) - \operatorname{E}(X+Y)\big)^2 \\
&= \operatorname{E}\big((X - \operatorname{E}X) + (Y - \operatorname{E}Y)\big)^2 \\
&= \operatorname{E}(X - \operatorname{E}X)^2 + 2\operatorname{E}(X - \operatorname{E}X)(Y - \operatorname{E}Y) \\
&\quad + \operatorname{E}(Y - \operatorname{E}Y)^2 \\
&= \operatorname{var}X + 2\operatorname{cov}(X,Y) + \operatorname{var}Y.
\end{aligned}
$$

Furthermore, if X and Y are independent, then $\operatorname{E}(XY) = \operatorname{E}X \operatorname{E}Y$, and hence that their covariance is zero. $\square$

Directly from the definition,

$$
\operatorname{var}(X) = \operatorname{cov}(X, X).
$$

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,

$$
\operatorname{var}(Y_1 + \cdots + Y_n) = n \operatorname{var}Y, \quad \operatorname{var}(nY) = n^2 \operatorname{var}Y.
$$

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:

<div align="center">

CORRELATION COEFFICIENT

</div>

The *correlation coefficient* of random variables X and Y whose variances are finite and non-zero is defined as

$$
\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}X}\sqrt{\operatorname{var}Y}}.
$$

As seen from theorem 10.2.33, the correlation coefficient of random variables equals the covariance of the standardized variables $\frac{1}{\sqrt{\operatorname{varX}}}(X - \operatorname{E}X)$ and $\frac{1}{\sqrt{\operatorname{varY}}}(Y - \operatorname{E}Y)$.

The following equalities hold (here, a, b, c, d are real constants, $bd \ne 0$, and X, Y are random variables with finite non-zero variances)

$$
\rho_{a+bX,c+dY} = \operatorname{sgn}(bd)\rho_{X,Y}
$$

$$
\rho_{X,X} = 1.
$$

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 $\operatorname{E}X$ with probability 1. If the value of X falls into an interval I not containing $\operatorname{E}X$ with probability $p \ne 0$, then the expression $\operatorname{var}X = \operatorname{E}(X - \operatorname{E}X)^2$ is positive. Stochastically, random variables with zero variance behave as constants.

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 $X \sim \text{Bi}(n, \frac{9}{10})$. By F, we have $\text{E}\,X = 0.9n$ and $\text{var}\,X = 0.09n$, so the standardized variable is $Z = \frac{X - 0.9n}{\sqrt{0.09n}}$. The condition in question can be written as

$$P(|X - 0.9n| \le 0.034n) = P\left(|Z| \le \frac{0.034n}{\sqrt{0.09n}}\right) =$$
$$= P\left(|Z| \le \frac{0.34}{3}\sqrt{n}\right) \ge 0.995.$$

By the de Moivre-Laplace theorem, for large n, the distribution function can be approximated by the distribution function Φ of the normal distribution. Thus,

$$P\left(|Z| \le \frac{0.34}{3}\sqrt{n}\right) \approx \Phi\left(\frac{0.34}{3}\sqrt{n}\right) - \Phi\left(-\frac{0.34}{3}\sqrt{n}\right) =$$
$$= 2\Phi\left(\frac{0.34}{3}\sqrt{n}\right) - 1.$$

Altogether, we get the condition

$$2\Phi\left(\frac{0.34}{3}\sqrt{n}\right) - 1 \ge 0.995.$$

Odtud vypočítáme $n \ge \left(\frac{3z(0.9975)}{0{,}34}\right)^2 \approx 615.$ $\qquad\square$

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 $\text{E}\,X_i = \frac{1}{\lambda}$ and $\text{var}\,X_i = \frac{1}{\lambda^2}$, respectively. Thus, the expected service life of each gadget is $\text{E}\,X_i = \mu = 10$ hours, with variance $\text{var}\,X_i = \sigma^2 = 100$ hours^2. 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

$$P(900 \le \sum X_i \le 1050) = P\left(-1 \le \frac{1}{100}\sum_{i=1}^{100} X_i - 10 \le 0{,}5\right)$$

can be approximated with the distribution function of the normal distribution:

If the covariance is a positive-definite symmetric bilinear form, then it would follow from the Cauchy-Schwarz inequality (see 3.4.3) that

$$(1) \qquad\qquad |\rho_{X,Y}| \le 1$$

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 a 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 $Z \sim \text{N}(0,1)$ and Z^2. Then $\text{cov}(Z, Z^2) = \text{E}\,Z^3 = 0$ 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* $|\rho_{X,Y}| \le 1$. *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 $Y + sX \sim \text{D}(c)$ 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:

$$0 \le \text{var}\left(\frac{Y - \text{E}\,Y}{\sqrt{\text{var}\,Y}} + t\,\frac{X - \text{E}\,X}{\sqrt{\text{var}\,X}}\right) = 1 + 2t\rho_{X,Y} + t^2.$$

The right-hand quadratic expression does not have two distinct real roots; hence its discriminant cannot be positive. So $4(\rho_{X,Y})^2 - 4 \le 0$. 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 asumes a fixed value with probability 1. This yields the affine relation as expected. $\qquad\square$

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.

$$P(900 \le \sum X_i \le 1050) \approx \Phi(0.5) - \Phi(-1) \approx 0.533.$$

$\square$

10.H.11. We keep putting items into a chest. The expected mass of an item is $3\,\mathrm{kg}$ and the standard deviation is $0.8\,\mathrm{kg}$. 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 $\mu = \mathrm{E}\,X_i = 3$ and $\sigma = \sqrt{\mathrm{var}\,X_i} = 0.8$ (in kilograms), and we want to have

$$P(\sum_{i=1}^{n} X_i \le 1000) = 0.99.$$

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

$$P(\sum_{i=1}^{n} X_i \le 1000) = P(S_n \le \frac{1000}{0.8\sqrt{n}} - \frac{3\sqrt{n}}{0.8}) \approx \Phi(\frac{1000}{0.8\sqrt{n}} - \frac{3\sqrt{n}}{0.8}).$$

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 μ of the normal distribution $\mathrm{N}(\mu, \sigma^2)$. 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, ∞). The formula for these intervals can be derived similarly as for the two-sided interval. Now, we have for the random variable $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathrm{N}(0,1)$ that

$$1 - \alpha = \Phi(z(1-\alpha)) = P(Z < z(1-\alpha)).$$

Hence it immediately follows that

$$1 - \alpha = P(\bar{X} - \frac{\sigma}{\sqrt{n}}z(1-\alpha) < \mu),$$

Consider a random vector $X = (X_1, \ldots, X_n)^T$ 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)

$$\mathrm{var}\,X = \mathrm{E}(X - \mathrm{E}\,X)(X - \mathrm{E}\,X)^T.$$

Using the definition of the expected value of a vector and expanding the matrix multiplication, it is immediate that the covariance matrix $\mathrm{var}\,X$ is the symmetric matrix

$$\begin{pmatrix} \mathrm{var}\,X_1 & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}\,X_2 & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}\,X_n \end{pmatrix}.$$

Theorem. *Consider a random vector $X = (X_1, \ldots, X_n)^T$ 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,*

$$\mathrm{var}(Y) = \mathrm{var}(BX + c) = B(\mathrm{var}\,X)B^T.$$

PROOF. The claim follows from direct computation, using the properties of the expected value:

$$\begin{aligned} \mathrm{var}(Y) &= \mathrm{E}\big((BX + c) - \mathrm{E}(BX + c) \big) \big((BX + c) \\ &\quad - \mathrm{E}(BX + c) \big)^T \\ &= \mathrm{E}(B(X - \mathrm{E}\,X))(B(X - \mathrm{E}\,X))^T \\ &= B\,\mathrm{E}(X - \mathrm{E}\,X)(X - \mathrm{E}\,X)^T B^T \\ &= B(\mathrm{var}\,X)B^T. \end{aligned}$$

$\square$

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 $\mu_k = E\big((X - EX)^k\big)$ 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 $\mathrm{E}\,|X|^k$.

From the definition it follows that for a continuous random variable X,

$$\mathrm{E}\,X^k = \int_{-\infty}^{\infty} x^k f_X(x)\,dx.$$

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 σ^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 χ^2. Then, we can immediately see that

$$1 - \alpha = P\left(\chi^2_{n-1}(\alpha/2) \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1}(1-\alpha/2)\right).$$

Thus, the two-sided $100(1-\alpha)\%$ confidence interval for the variance is

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)}\right)$$

and similarly for the one-sided upper and lower estimates, we get

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha)}, \text{ resp. } \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha)} \leq \sigma^2.$$

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 $\text{N}(100, \frac{250}{3})$. 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 μ 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 $\text{N}(\mu, \sigma^2)$ with unknown expected value μ 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 μ,

ii) the lower estimate for μ at significance level 95 %.

Similarly, for a discrete random variable X whose probability is concentrated into points x_i,

$$\text{E}\, X^k = \sum_i x_i^k f_X(x_i).$$

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 $M_X(t) : \mathbb{R} \to \mathbb{R}$ defined by

$$M_X(t) = \text{E}\, e^{tX} = \begin{cases} \sum_i e^{tx_i} f_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

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*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\, \text{E}\, X^k.$$

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

$$\frac{d^k}{dt^k}M_X(t) = \text{E}\, X^k,$$

as expected.

The second claim is obvious for two discrete variables X and Y with only a finite number of values $x_1, \ldots, x_k$ 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

$$M(t) = e^{tx_1} f(x_i) + \cdots + e^{tx_k} f(x_k)$$

must be the shared probability function values for both random variables X and Y.

Solution. a) By 10.3.4, the $100(1 - \alpha)\%$ two-sided confidence interval for the unknown expected value μ 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, σ^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 μ 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 $\mathrm{N}(10\,\mathrm{mm}; 0.0734\,\mathrm{mm}^2)$. 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 χ^2_{n-1}. In our case, $n = 21$ and $\sigma^2 = 0.0734$, so

$$P(S \leq 0.2) = P \left(\frac{20}{0.0734} S^2 \leq \frac{20}{0.0734} 0.2^2 \right) = \chi^2_{20} \left(\frac{20 \cdot 0.2^2}{0.0734} \right).$$

The expression in the argument of the distribution function is approximately 10.9, and we can learn from the table of the χ^2 distribution that $\chi^2_{20}(10.9) \approx 0.05$. Thus, the probability that delivery will be accepted is only 5 %. We could have expected the probability to be low: indeed, $\mathrm{E}\, S^2 = \sigma^2 = 0.0734 > 0.2^2$. If the customer tests only 4 products, then the probability of acceptance is given by the expression $\chi^2_3 \left(\frac{3 \cdot 0.2^2}{0.0734} \right) \approx \chi^2_3(1.63)$. The value of the distribution function of χ^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 $\chi^2_3(0.58) = 0.1$ and $\chi^2_3(6.25) = 0.9$, then

$$\chi^2_3(1.63) \approx (1.63 - 0.58) \frac{0.9 - 0.1}{6.25 - 0.58} + 0.1 \approx 0.24.$$

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,

$$M(it) = \int_{-\infty}^{\infty} \mathrm{e}^{itx} f(x) dx,$$

which is the inverse Fourier transform of $f(x)$, up to the constant multiple $\sqrt{2\pi}$, see 7.2.5 (on page 474). 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*

$$M_{a+bX}(t) = \mathrm{e}^{at} M_X(bt)$$
$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

PROOF. The first formula can be computed directly from the definition:

$$M_V(t) = \mathrm{E}\, \mathrm{e}^{(a+bX)t} = \mathrm{E}\, \mathrm{e}^{at}\, \mathrm{e}^{(bt)X} = \mathrm{e}^{at} M_X(bt).$$

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.

$$M_W(t) = \mathrm{E}\, \mathrm{e}^{t(X+Y)} = \mathrm{E}\, \mathrm{e}^{tX}\, \mathrm{e}^{tY}$$
$$= \mathrm{E}\, \mathrm{e}^{tX}\, \mathrm{E}\, \mathrm{e}^{tY} = M_X(t) M_Y(t). \qquad \square$$

10.2.38. Normal and binomial distributions. As an illustrating example, compute the moment function of two random variables $X \sim \mathrm{N}(\mu, \sigma)$ and $X \sim \mathrm{Bi}(n, p)$.

MOMENT GENERATING FUNCTION FOR $\mathrm{N}(\mu, \sigma)$

Proposition. *If $X \sim \mathrm{N}(\mu, \sigma)$, then*

$$M_X(t) = \mathrm{e}^{\mu t}\, \mathrm{e}^{\frac{\sigma^2 t^2}{2}}.$$

In particular, it is an analytic function on all of $\mathbb{R}$.

Although this results is only an estimate, we can be sure that the probability of acceptance is much greater than when testing 21 products. □

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 $(1.22, 1.58)$. □

10.I.5. Let $X_1, \ldots, X_n$ 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 μ 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

$$\frac{2 \cdot 0.2}{\sqrt{n}} z(0.975) \le 0.16.$$

Since $z(0.975) \approx 1.96$, we obtain $n \ge 24.01$. Thus, at least 25 measurements are necessary. □

10.I.6. Consider a random variable X with distribution $N(\mu, \sigma^2)$, where μ, σ^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 μ.

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 μ, when the variance

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:

$$
\begin{aligned}
M_Z(t) &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{t^2}{2} \right)^k = \sum_{k=0}^{\infty} \frac{1}{k! 2^k} t^{2k} = \\
&= 1 + 0\,t + \frac{1}{2}\,t^2 + 0\,t^3 + \frac{3}{4!}\,t^4 + \ldots
\end{aligned}
$$

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 $X \sim Bi(n, p)$,

$$
\begin{aligned}
M_X(t) = E\,e^{tX} &= \sum_{k=0}^{n} (p\,e^t)^k \binom{n}{k} (1-p)^{n-k} \\
&= (p\,e^t + (1-p))^n = (p(e^t - 1) + 1)^n \\
&= 1 + npt + \frac{1}{2}\big(n(n-1)p^2 + np\big)t^2 + \ldots
\end{aligned}
$$

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 $\text{var}\,X = E\,X^2 - (E\,X)^2 = np(1-p)$.

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 $\mu \in (14.0, 16.7)$. □

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
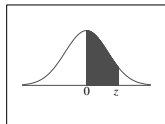
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.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

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Solution. Let X denote the random variable that corresponds to the number of heads. Then, X has binomial distribution $\mathrm{Bi}(3600, 1/2)$ (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 Φ of the normal distribution. Thus, the wanted probability is

$$P = 1 - P[1755 \le X \le 1845] =$$
$$= 1 - P\left[-1.5 \le \frac{X-1800}{30} \le 1.5 \right] = 2\Phi(-1.5) \doteq 0.1336,$$

where the last value was learned from the table. □

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

$$\gamma_1 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^3}{(\sqrt{\mathrm{var}\,X})^3},$$

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

$$\gamma_2 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^4}{(\mathrm{var}\,X)^2} - 3.$$

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 $X_n \sim \mathrm{Bi}(n,p)$, can be viewed as the arithmetic means of n independent variables with distribution $\mathrm{A}(p)$, 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 $\mathrm{cov}(X_i, X_j) = 0$ guarantees that the variances sum up.

THE LAW OF LARGE NUMBERS

Proposition. *Consider a sequence of pairwise uncorrelated random variables $X_1, X_2, \ldots$ which have the same finite expected value $\mathrm{E}\,X_i = \mu$. Moreover, assume the variances are bounded, so that $\mathrm{var}\,X_i \le C$, for a fixed constant C. Then for any $\varepsilon > 0$,*

$$\lim_{n \to \infty} P\big(\big| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \big| < \varepsilon \big) = 1.$$

PROOF. By the use Chebyshev's inequality just as at the end of subsection 10.2.32,

$$P\big(\big| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \big| \ge \varepsilon \big) \le \frac{\mathrm{var}\big(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \big)}{\varepsilon^2}$$
$$= \frac{\frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}\,X_i}{\varepsilon^2} \le \frac{C}{n\varepsilon^2}.$$

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.

$$P[X < 5000] = P[\underbrace{\frac{X - 5150}{\sqrt{5150 \cdot 0.485}}}_{\sim N(0,1)} < \underbrace{\frac{-150}{\sqrt{5150 \cdot 0.485}}}_{-3,001...}] \doteq$$

$$\doteq 0.00135$$

$\square$

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 Φ 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, $i = 1 \ldots 5$ 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

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. $\square$

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 μ 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 $\mathrm{E}\, X_i = 0$ and $\mathrm{var}\, X_i = 1$. 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)

$$M_{S_n}(t) = \mathrm{E}\, e^{\frac{t}{\sqrt{n}}\sum_i X_i} = \left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Since it is assumed that the variables X_i are standardized,

$$M_X\left(\frac{t}{\sqrt{n}}\right) = 1 + 0\,\frac{t}{\sqrt{n}} + 1\,\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right),$$

where again $o(G(n))$ is written for expressions which, when divided by $G(n)$, approach zero as $n \to \infty$, see subsection 6.1.16.

Thus, in the limit,

$$\lim_{n\to\infty} M_{S_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = e^{\frac{t^2}{2}}.$$

This is just the moment generating function of the normal distribution $Z \sim \mathrm{N}(0,1)$, 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!

rearrange the expression:

$$P\left[\left|\frac{X}{n} - p\right| < 0.02\right]$$

$$= P\left[-0.02 < \frac{X}{n} - p < 0.02\right] =$$

$$= P\left[-0.02 \cdot n < X - pn < 0.02 \cdot n\right] =$$

$$= P\left[\frac{-0.02 \cdot n}{\sqrt{np(1-p)}} < \frac{X - pn}{\sqrt{np(1-p)}} < \frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right] =$$

$$= \Phi\left(\frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right) =$$

$$= 2\Phi\left(\frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right) - 1,$$

where Φ is the distribution function of the normal distribution. Thus, let us solve the inequality

$$2\Phi\left(\frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right) - 1 \geq 0.99$$

$$\Phi\left(\frac{0.02 \cdot n}{\sqrt{np(1-p)}}\right) \geq 0.995$$

Since the distribution function is increasing, the last condition is equivalent to

$$\frac{0.02 \cdot n}{\sqrt{np(1-p)}} \geq \Phi^{-1}(0.995)$$

$$\frac{0.02 \cdot n}{\sqrt{np(1-p)}} \geq 2.576$$

$$\sqrt{n} \geq 50 \cdot 2.576 \cdot \underbrace{\sqrt{p(1-p)}}_{\leq \frac{1}{2}} \implies$$

$$\implies n \geq (25 \cdot 2.276)^2 \cdot 4147$$

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).

$\square$

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 μ_i, and the variance σ is the same for all the components of both vectors.

Theorem. *Consider a sequence of independent random variables X_i which have the same expected value $\mathrm{E}\, X_i = \mu$, variance $\mathrm{var}\, X_i = \sigma^2 > 0$ and uniformly bounded third absolute moment $\mathrm{E}\, |X_i|^3 < C$. 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

$$\lim_{n \to \infty} P(S_n < x) = \Phi(x),$$

where Φ 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 $\mathrm{E}\, X_i = 0$ and $\mathrm{var}\, X_i = 1$ 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 $\mathrm{E}\, |X_i|^3$ 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 μ_i and variances σ_i^2. Write

$$s_n = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

and assume for some $\delta > 0$

$$\lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\, |X_i - \mu_i|^{2+\delta} = 0.$$

Then $\frac{X_i - \mu_i}{s_n}$ converges in distribution to $Z \sim \mathrm{N}(0, 1)$.

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

$$0 \leq \lim_{n \to \infty} n^{-3/2} \sigma^{-3} \sum_{i=1}^{n} \mathrm{E}\, |X_i|^3 \leq C\sigma^{-3} \lim_{n \to \infty} n^{-3/2+1} = 0.$$

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 $\mathrm{Bi}(n, p)$. 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 $\mathrm{E}\, |X_i|^3 = p < 1$.

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 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 $\mathrm{t}_{n_1+n_2-2}$, so the null hypothesis $\mu_1 = \mu_2$ is rejected at level α if we have

$$
|T| \geq \mathrm{t}_{n_1+n_2-2}(\alpha). \qquad \square
$$

10.I.12. In JZD[1] 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.

a) 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.

b) On the 5% level, test the hypothesis that both farms have cows of the same quality.

---

[1] JZD — jednotné zemědělské družstvo — an agricultural cooperative farm, created by forced collectivization in 1950s in Czechoslovakia.

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 $X \sim \mathrm{Bi}(n,p)$ behaves as the random variable with normal distribution $\mathrm{N}(np, np(1-p))$ 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 $X \sim \mathrm{Bi}(n,p)$. 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

$$
0.9 \simeq P\left(\left| \frac{1}{n}X - p \right| < 0.05 \right)
$$

$$
= P\left(-\frac{0.05n}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{0.05n}{\sqrt{np(1-p)}} \right)
$$

$$
\simeq \Phi\left(\frac{0.05n}{\sqrt{np(1-p)}} \right) - \Phi\left(-\frac{0.05n}{\sqrt{np(1-p)}} \right)
$$

$$
= 2\Phi\left(\frac{0.05n}{\sqrt{np(1-p)}} \right) - 1,
$$

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.

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 $\sigma = 2 \ hl$.

Solution. First of all, let us compute the results for the known variance. In order to find the confidence interval, we use the statistic

$$U = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}},$$

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

$$\overline{X}_1 = \frac{15 + 14 + 13 + 16 + 17}{5} = 15,$$

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

$$\overline{X}_2 = \frac{12 + 16 + 13 + 15 + 13 + 11 + 18}{7} = 14,$$

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

$$T = \frac{\overline{X} - \mu}{S \sqrt{n}},$$

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

$$(13.03; 16.97).$$

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 $x \le 0$, while for positive x,

$$F_Y(x) = P(Y < x) = P(-\sqrt{x} < Z < \sqrt{x})$$
$$= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = \int_0^x \frac{1}{\sqrt{2\pi}} t^{-1/2} e^{-t/2} \, dt.$$

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 χ^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

$$f_X(x) = c x^{a-1} e^{-bx}$$

for $x > 0$, while $f_X(x) = 0$ for non-positive x, i.e., the distribution χ^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:

$$E\, X^k = \int_0^\infty x^k \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \, dx$$
$$= \frac{\Gamma(a+k)}{\Gamma(a) b^k} \int_0^\infty \frac{b^{a+k}}{\Gamma(a+k)} x^{a-1+k} e^{-bx} \, dx$$
$$= \frac{\Gamma(a+k)}{\Gamma(a) b^k},$$

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, $-b < t < b$

$$M_X(t) = \int_0^\infty e^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \, dx =$$
$$= \frac{b^a}{(b-t)^a} \int_0^\infty x^k \frac{(b-t)^a}{\Gamma(a)} x^{a-1} e^{-(b-t)x} \, dx =$$
$$= \frac{b^a}{(b-t)^a}.$$

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.

Thus, let us examine the hypothesis assuming the known variances $\sigma_1^2 = \sigma_2^2 = 4$. We use the statistic

$$
\begin{aligned}
U &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \\
&= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1),
\end{aligned}
$$

where μ_1 and μ_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

$$
\begin{aligned}
K &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \\
&= \frac{\overline{X}_1 - \overline{X}_2}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2},
\end{aligned}
$$

where

$$S_* = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Thus, for the sum of independent variables $Y = X_1 + \cdots + X_n$ with distributions $X_i \sim \Gamma(a_i, b)$, the moment generating function (for values $|t| < b$) is obtained

$$M_Y(t) = \left(\frac{b}{b - t}\right)^{a_1 + \cdots + a_n},$$

that is, $Y \sim \Gamma(a_1 + \cdots + a_n, b)$. It is essential that all of the gamma distributions share the same value of b.

As an immediate corollary, the density of the variable $Y = Z_1^2 + \cdots + Z_n^2$ 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 χ^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 $f_Y(x)$ 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

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.

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 μ_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 μ_i, while in the submodel, there is a single parameter μ, so the

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)

$$
F_Y(y) = P(\sqrt{X} < y) = P(X < y^2)
$$
$$
= \int_0^{y^2} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} dx
$$
$$
= \int_0^{y} \frac{1}{2^{n/2-1}\Gamma(n/2)} t^{n-1} e^{-t^2/2} dt.
$$

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 $Z = (Z_1, \ldots, Z_n)$ with independent components $Z_i \sim N(0,1)$. Then its covariance matrix is equal to the unit matrix, i.e., $\mathrm{var}\, Z = \mathbb{I}_n$.

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 $\mathrm{E}\, U = a$ and covariance matrix $\mathrm{var}\, U = V = BB^T$ (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* $\mathrm{N}_m(a, V)$.

tested statistic is of the form

$$F = \frac{(n-k)}{(k-1)} \frac{(RSS^0 - RSS)}{RSS}.$$

$\square$

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.

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 $b_1 X + b_0$ 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 $(1, \ldots, 1)$ and $(x_1, \ldots, x_n)$. For parameters b_0, b_1 of the regression line $Y = b_1 X + b_0$, we obtain

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$
$$= \frac{(1 - 5.4)(3 - 9.6) + \cdots + (10 - 5.4)(18 - 9.6)}{((1 - 5.4)^2 + (4 - 5.4)^2 + (5 - 5.4)^2 + (7 - 5.4)^2 + (10 - 5.4)^2} =$$
$$= 1.677.$$

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:

$$X = 0.5867 \cdot Y - 0.2322.$$

$\square$

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 $z_i < u_i$)

$$F_U(u) = P(U_i < u_i, i = 1, \ldots, n) =$$
$$= \int \cdots \int_{Q^T z < u} (2\pi)^{-n/2} e^{-\sum z_i^2 / 2} dz_1 \cdots dz_n$$
$$= \int \cdots \int_{t < u} (2\pi)^{-n/2} e^{-\sum t_i^2 / 2} dt_1 \cdots dt_n$$
$$= \left(\int_{-\infty}^{u_1} (2\pi)^{-1/2} e^{-t_1^2} dt_1 \right) \cdots$$
$$\cdots \left(\int_{-\infty}^{u_n} (2\pi)^{-1/2} e^{-t_n^2} dt_n \right)$$
$$= F_{U_1}(u_1) \cdots F_{U_n}(u_n)$$

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).

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 $v = (v_1, v_2, \ldots, v_n)$ dependent on $x = (x_1, \ldots, x_n)$ and $x^2 = (x_1^2, \ldots, x_n^2)$. 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 $b_2 x^2 + b_1 x + b_0$ would 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

$$v = 0.29 x^2 + 0.47 x + 9.26,$$

so the estimated velocity on the tenth day is approximately 42.96 km/s. In the model of classic linear regression, we would get

$$v = 2.18 x + 7.26,$$

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.

$\square$

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 θ 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 θ. 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

K. Bayesian data analysis

10.K.1. Consider the Bernoulli process defined by a random variable $X \sim \mathrm{Bi}(n, \theta)$ with binomial distribution, and assume that the parameter θ 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 γ?

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 θ and γ to compute $\theta = \frac{\gamma}{1+\gamma}$. In addition, we can immediately see that the probability density of γ is non-zero only for positive values of the variable. Now, we can formulate the statement as the requirement

$$(1) \qquad \Theta = P(\theta < \Theta) = P\left(\gamma < \frac{\Gamma}{1+\Gamma}\right) = \int_0^\Gamma f(s)\, ds,$$

where $\Gamma = \frac{\Theta}{1-\Theta}$. However, the right-hand upper bound contains the changing limit γ, 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. □

We could see in subsection 10.3.7 that when taking the Bayesian approach with binomial model of probability distribution of a random variable $X \sim \mathrm{Bi}(n, \theta)$, 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 θ on the interval $(0, 1)$. Thus, it is just the a posteriori probability distribution of θ 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 charakteristics of the so-called *beta-distribution* 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 $\mathrm{B}(a, b)$, known as *beta-function* in mathematical analysis and other sciences (e. g. physics). The function gamma, which generalizes the discrete values of factorial,

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 $(x_1, \ldots, x_N)$. From this set there is only a *sample* with values $(X_1, \ldots, X_n)$.

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 $(X_1, \ldots, X_n)$ of independent, identically distributed random variables. In particular, they have the same distribution function $F_X(x)$ and moments

$$\mathrm{E}\, X_i = \mu, \quad \mathrm{var}\, X_i = \sigma^2.$$

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 μ and finite variance σ^2 satisfies*

$$\mathrm{E}\, \bar{X} = \mu, \quad \mathrm{var}\, \bar{X} = \frac{1}{n}\sigma^2.$$

The sample variance S^2 satisfies

$$\mathrm{E}\, S^2 = \sigma^2.$$

PROOF. As derived in subsection 10.2.29,

$$\mathrm{E}\, \bar{X} = \frac{1}{n}\, \mathrm{E} \sum_{i=1}^n X_i = \frac{1}{n} n\mu = \mu.$$

emerges in the following calculation:

$$\Gamma(x)\Gamma(y) = \int_0^\infty \mathrm{e}^{-t}\, t^{x-1} dt \cdot \int_0^\infty \mathrm{e}^{-s}\, s^{y-1} ds =$$

$$= \int_0^\infty \int_0^\infty \mathrm{e}^{-t-s}\, t^{x-1} s^{y-1} dt\, ds =$$

(substitution $t = rq$, $s = r(1-q)$)

$$= \int_{r=0}^\infty \int_{q=0}^1 \mathrm{e}^{-r}(rq)^{x-1}\big(r(1-q)\big)^{y-1} r\, dq\, dr =$$

$$= \int_{r=0}^\infty \mathrm{e}^{-r}\, r^{x+y-1} dr \cdot \int_{t=0}^1 q^{x-1}(1-q)^{y-1} dq =$$

$$= \Gamma(x+y)\mathrm{B}(x,y).$$

Thus, we get the general formula

$$\mathrm{B}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

and it follows from properties of the gamma-function for positive integers a, b that

$$\mathrm{B}(n-k+1, k+1) = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}\binom{n}{k}^{-1}.$$

We can directly compute that the expected value of the variable $X \sim \mathrm{fi}(a,b)$ with beta-distribution is (applying $\Gamma(z+1) = z\Gamma(z)$)

$$\mathrm{E}\,X = \frac{\mathrm{B}(a+1,b)}{\mathrm{B}(a,b)} = \frac{a}{a+b}.$$

If $a = b$, then the expected value and median are $\frac{1}{2}$.

We can also directly calculate the variance

$$\mathrm{var}\,X = \mathrm{E}(X - \mathrm{E}\,X)^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

Thus, for $a = b$, we get $\mathrm{var}\,X = \frac{1}{8a+4}$, which shows that the variance decreases as $a = b$ increases. For $a = b = 1$, we get the ordinary uniform distribution on the interval $(0,1)$. □

10.K.3. In the situation as in the problem above the previous one, assume that the success chance θ in the Bernoulli process is a random variable with probability distribution $\mathrm{fi}(a,b)$. 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 discusses the special case with uniform distribution $\mathrm{fi}(1,1)$. Thus, we can continue with the equality $\|1\|$, where we used the form of this distribution. Now the left-hand side contains, instead of Θ, the expression

$$\frac{1}{\mathrm{B}(a,b)} \int_0^\Theta t^{a-1}(1-t)^{b-1} dt.$$

When differentiating, we must use the rule for differentiation of integral with variable upper bound. Thus, we get for the

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

$$\mathrm{var}\,\bar{X} = \frac{1}{n^2} \mathrm{var} \sum_{i=1}^n X_i = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2.$$

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:

$$\mathrm{E}\,s^2 = \frac{1}{n}\mathrm{E}\sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{n}(\bar{X} - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^n \mathrm{var}\,X_i - \mathrm{var}\,\bar{X} =$$

$$= \big(1 - \frac{1}{n}\big)\sigma^2.$$

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 σ. Of course, this multiplication makes sense only if $n \neq 1$. □

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, $\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.

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 $(\sqrt{n})^{-1}(1,\ldots,1)$. Then $U_1^2 = n\bar{Z}^2$, so we

wanted density that

$$\mathrm{B}(a,b)f(s) = \left(\frac{s}{s+1}\right)^{a-1}\left(1-\frac{s}{s+1}\right)^{b-1}\frac{1}{(s+1)^2} =$$

$$= \left(\frac{s^{a-1}}{s+1}\right)^{a+b}.$$

The picture shows the densities for $a = b = p = 2, 5, 15$.
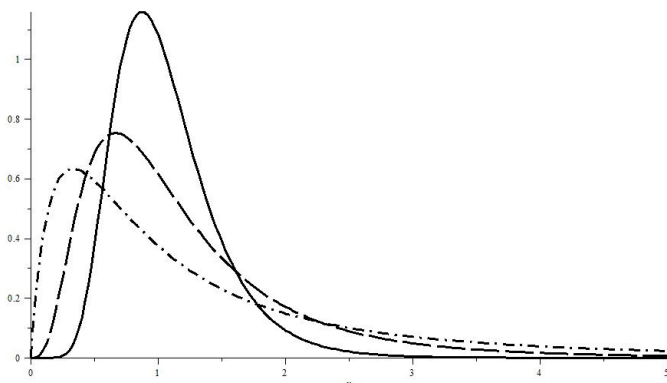
This enforces the intuition that the same and not too small values of $a = b = p$ correspond to the most probable value $\theta = \frac{1}{2}$, so the density of the chance is greatest around one. The higher p, the lower the variance of this variable. $\qquad\square$

10.K.4. Show that the Bernoulli experiment, described by a random variable $X \sim \mathrm{Bi}(n,\theta)$, and the a priori probability of a random variable θ 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 θ (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

$$g(\theta) = \frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

and the probability of the examined variable X provided the value of θ occurred. Thus, assuming k successes in the Bernoulli experiment, we get the a posteriori density (the sign used instead of equality denotes "proportional")

$$g(\theta|X=k) \propto P(X=k|\theta)g(\theta) \propto$$
$$\propto \theta^k(1-\theta)^{n-k}\theta^{a-1}(1-\theta)^{b-1} =$$
$$= \theta^{a+k-1}(1-\theta)^{b+n-k-1}.$$

can compute:

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 + n\bar{Z}^2$$

$$\sum_{i=2}^{n} U_i^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

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 $(X_1,\ldots,X_n)$ be a random sample from the $\mathrm{N}(\mu,\sigma^2)$ distribution. Then, $\bar{X}$ and S^2 are independent variables, and*

$$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2), \quad \frac{n-1}{\sigma^2}S^2 \sim \chi_{n-1}^2.$$

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 everything 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 $\mathrm{N}(\mu,\sigma^2)$. 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, $\bar{X} \sim \mathrm{N}(\mu,\sigma^2/n)$, so $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathrm{N}(0,1)$. 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 α, $0 < \alpha < 1$. Consider μ to be the unknown parameter, while the variance can be assumed (be it correct or not) to remain unchanged. It follows that

$$1-\alpha = P(|Z| < z(\alpha/2)) = P\left(\left|\sqrt{n}\frac{\bar{X}-\mu}{\sigma}\right| < z(\alpha/2)\right)$$

$$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right),$$

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 μ with an a priori fixed probability. The middle point of this interval is called the *point estimate* for parameter μ; the whole interval is called the *interval estimate*. We can also say that at the confidence

Thus, we have indeed obtained the density (up to a constant, which we need not evaluate) of the a posteriori distribution for θ with distribution $\mathrm{B}(a + k, b + n - k)$.

Its a posteriori expected value is

$$\hat{\theta} = \frac{a + k}{a + b + n}.$$

For n and k approaching infinity so that $k/n \to p$, our a posteriori estimate also satisfies $\hat{\theta} \to p$. 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. □

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, $j = 1, \ldots, N$, 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). [2]

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 $\mathrm{Bi}(n, p_j)$.

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.

level α, the estimated parameter μ is or is not different from another value μ_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 $T \sim \mathrm{t}_{n-1}$, 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

$$\mu \in \big(4.412,\ 5.854\big), \quad \mu \in \big(4.321,\ 5.945\big).$$

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 θ (a vector in general).

We are trying to find a *sample statistic* T, i.e., a function of the random variables $X_1, X_2, \ldots$ which, in a mathematical sense, estimates the actual value of the parameter θ. T is said to be an *unbiased estimator* of θ if and only if $\mathrm{E}\, T = \theta$. The expected value $\mathrm{E}(T - \theta)$ 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

---

[2]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 realistic method is to use the same assumption for the a priori distribution of the probabilities p_j of accident rates of individual drivers. In practice, one often uses a model with the Poisson distribution $\mathrm{Po}(\lambda_j)$ for the j-th driver, with further assumptions about the distribution of the parameter λ among the drivers. We can also assume quite well (and simply) that the distribution is $p_j \sim \mathrm{fi}(a, b)$ 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 $(p_j|S_j = k) = \mathrm{fi}(a + k, b + n - k)$, 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

$$\mathrm{E}\,X_{ji} = \mathrm{E}\,\mathrm{E}(X_{ji}|p) = \mathrm{E}\,p = p_0$$
$$\frac{\mathrm{E}\,\mathrm{var}(X_{ji}|p)}{\mathrm{var}\,\mathrm{E}(X_{ji}|p)} = \frac{\mathrm{E}(p(1 - p))}{\mathrm{var}\,p} = a + b = n_0$$

and the left-hand variables can be estimated directly.

$$\mathrm{E}\,X_{ij} = \mathrm{E}\,\mathrm{E}(X_{ji}|p) \simeq \frac{1}{N}\sum_{j=1}^{N}\hat{p}_j$$

$$\mathrm{E}\,\mathrm{var}(X_{ji}|p) \simeq \frac{1}{N}\sum_{j=1}^{N}\left(\frac{n}{n - 1}\hat{p}_j(1 - \hat{p}_j)\right)$$

$$\mathrm{var}\,\mathrm{E}(X_{ji}|p) \simeq s_{\hat{p}_j}^2 - \frac{1}{nN}\sum_{j=1}^{N}\left(\frac{n}{n - 1}\hat{p}_j(1 - \hat{p}_j)\right),$$

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

$$\hat{p}_j^b = 0.154 \cdot 0.145 + 0.846 \cdot \hat{p}_j.$$

said to be a *consistent estimator* of the parameter θ if and only if $T(n)$ converges in probability to θ, i.e., for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|T(n) - \theta| < \varepsilon) = 1.$$

The Chebyshev's inequality immediately yields

$$P(|T(n) - \mathrm{E}\,T_n| < \varepsilon) \geq 1 - \frac{\mathrm{var}\,T(n)}{\varepsilon^2}.$$

Assuming $\lim_{n\to\infty} \mathrm{E}\,T(n) = \theta$, then, for sufficiently large values of n,

$$P(|T(n) - \theta| < 2\varepsilon) \geq P(|T(n) - \mathrm{E}\,T(n)| < \varepsilon) \geq 1 - \frac{\mathrm{var}\,T(n)}{\varepsilon^2}.$$

A useful proposition is thus proved :

Theorem. *Assume that* $\lim_{n\to\infty} \mathrm{E}\,T(n) = \theta$ *and* $\lim_{n\to\infty} \mathrm{var}\,T(n) = 0$. *Then,* $T(n)$ *is a consistent estimator of* θ.

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, $\lim_{n\to\infty} \hat{\sigma}^2 = \sigma^2$, and it can be calculated that

$$\lim_{n\to\infty} \mathrm{var}\,\hat{\sigma}^2 = \lim_{n\to\infty} \mathrm{var}\,S^2 = \lim_{n\to\infty} \frac{2\sigma}{n - 1} = 0.$$

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 θ of random variables X_i. The value X_1 is, of course, an unbiased estimator of θ 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 θ (a vector in general). By the assumed independence, the joint density of the vector $(X_1, \ldots, X_n)$ is equal to the product:

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \theta) \cdots f(x_n, \theta),$$

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.

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. $\qquad\square$

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	18
total seeds	36	39	19	97	37	26	45	53	64	62

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

$$K = \sum_{j=1}^{r} \frac{(n_j - np_j)}{np_j},$$

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, $p_j = 0.75$, $j = 1, \ldots, 10$. 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.

Usually, it is more efficient to work with the *log-likelihood function*

$$\ell(x_1, \ldots, x_n, \theta) = \ln f(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ln f(x_i, \theta).$$

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 $f(x_1, \ldots, x_n, \theta) = 0$, set $\ell(x_1, \ldots, x_n, \theta) = -\infty$.

In the case of discrete random variables, use the same definition with probability mass function instead of the density, i.e.,

$$\ell(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ln(P(X_i = x_i | \theta)).$$

We can illustrate the principle on a random sample from the normal distribution $N(\mu, \sigma^2)$ with size n. The unknown parameters are μ or σ, 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 σ^2 is treated as a symbol for a variable):

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (-2)(x_i - \mu) = \frac{1}{\sigma^2} \left(-n\mu + \sum_{i=1}^{n} x_i\right)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 =$$

$$= \frac{1}{2(\sigma^2)^2} \left(-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

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 ℓ:

$$\begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}.$$

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 μ and σ, 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 $X_1, \ldots, X_{15}$ (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 μ (the expected value of the points of satisfaction), is viewed as the random variable

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
⋮	⋮	⋮	⋮	⋮
10	44	0.75	46.5	0.134409

The value of the statistic K for the given data is

$$K = 0.148148 + 0.258547 + \cdots + 0.134409 = 1.797495.$$

This value is less than $\chi^2_{0.95}(9) = 16.9$, so we do not reject the null hypothesis at level 0.05 (i. e., we do not refute the known genetic model).

□

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, Θ) has joint density $f(x, \theta)$, then the conditional probability of a component Θ, 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 θ 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 θ and thus the proper density is obtained from $f(x|\theta)g(\theta)$ by multiplying with a uniquely given constent 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 θ 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 θ),

$$g(\theta|x) \propto f(x|\theta)g(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-a)^2}{2b^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\theta^2\left(\frac{1}{\sigma^2} + \frac{1}{b^2}\right) - 2\theta\left(\frac{x}{\sigma^2} + \frac{a}{b^2}\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\theta - \frac{b^2 x + \sigma^2 a}{\sigma^2 b^2}\frac{\sigma^2 b^2}{b^2 + \sigma^2}\right)^2\left(\frac{b^2 \sigma^2}{b^2 + \sigma^2}\right)^{-1}\right).$$

This proves already that the distribution for θ 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, σ 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 σ 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 $\sigma \to 0$, 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 σ 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 σ^2/n is substituted instead of σ^2. To facilitate the notation, define the constant

$$c_n = \frac{nb^2}{nb^2 + \sigma^2}.$$

The a posteriori estimate for θ based on the found sample mean $\bar{X}$ has the distribution with parameters

$$\theta \sim \mathrm{N}(c_n \bar{X} + (1 - c_n)a, c_n \sigma^2/n).$$

As could be expected, for increasing n, the expected value of the distribution for θ 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

$$\theta \sim \mathrm{N}(5.230,\ 0.256).$$

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 θ 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 $X = (X_1, \ldots, X_n)$ (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 α. This means that a fixed bound for the probability of the type I error is required: the *significance level* α. 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 θ, 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 θ_0 and θ_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)

$$D \geq c \int_{W_c \setminus W} f(x, \theta_0) - c \int_{W \setminus W_c} f(x, \theta_0) =$$

$$= c \int_{W_c} f(x, \theta_0) - c \int_W f(x, \theta_0) = c\alpha - c\alpha = 0.$$

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 μ_0", while H_A says that it is equal to a different value μ_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 μ_1 is not essential. In the context of classical probability, the decision at a given level α whether or not there is a change to the expected value μ 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

$$\frac{f(x, \mu_1, \sigma^2)}{f(x, \mu_0, \sigma^2)} = e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i - \mu_1)^2 - (x_i - \mu_0)^2\right)} \geq c.$$

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 α, 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 μ_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 $z = \frac{5.133 - 6}{4}\sqrt{15} = -1.678$, while $-z(0.05) = -1.645$.

Therefore, if we are testing whether the new teacher is even worse 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 $Y = (Y_1, \ldots, Y_n)^T$ 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, β 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 β 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 β, 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 $\operatorname{span} X \subset \mathbb{R}^n$ generated by the columns of the matrix X. This is minimizing the function

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\Big)^2.$$

Choose an arbitrary orthonormal basis of the vector subspace $\operatorname{span} X$ 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 $\operatorname{span} X$, 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,

$$(1) \qquad b = (X^T X)^{-1} X^T Y.$$

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):

$$b = (T^T P^T PT)^{-1} T^T P^T Y =$$
$$= T^{-1} (T^T)^{-1} T^T P^T (PT\beta + \sigma Z) =$$
$$= \beta + \sigma T^{-1} V.$$

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 $\hat{Y}$,

$$\hat{Y} = X\beta + \sigma PV, \quad \hat{Y} \sim \mathrm{N}(X\beta, \sigma^2 PP^T).$$

(2) The residual sum of squares and the normed square of the residue size have distributions:

$$Y - \hat{Y} \sim \mathrm{N}(0, \sigma^2 RR^T), \quad \|Y - \hat{Y}\|^2/\sigma^2 \sim \chi^2_{n-k}.$$

(3) The random variable $b = \beta + \sigma T^{-1} V$ has distribution

$$b \sim \mathrm{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

(4) The residual variance satisfies $(n - k)S^2/\sigma^2 \sim \chi^2_{n-k}$.

(5) The expected value of the residual variance is $\mathrm{E}\, S^2 = \sigma^2$.

(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 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 $\mathrm{N}(0,1)$, so it is the distribution χ^2_{n-k}, 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 χ^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. $\qquad\square$

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 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 generates 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:

$$\hat{Y}^0 = P^0 (P^0)^T Y = X^0 \beta^0 + \sigma P^0 V^0$$

$$Y - \hat{Y}^0 = \sigma P^1 V^1 + \sigma RU$$

$$\|Y - \hat{Y}^0\|^2 = \sigma^2 \|V^1\|^2 + \sigma^2 \|U\|^2$$

$$(\mathrm{RSS}^0 - \mathrm{RSS})/\sigma^2 = \|V^1\|^2.$$

Therefore, the normed difference of the residues has distribution χ^2_{k-q}. It follows immediately that the statistic F given as the relative difference of the residues has Fisher-Snedecor distribution:

$$F = \frac{(\mathrm{RSS}^0 - \mathrm{RSS})/(k-q)}{\mathrm{RSS}/(n-k)} \sim \mathrm{F}_{k-q,n-k}\,.$$

In practice, the parameter σ 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 β equals a given value β_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 α 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 , 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, $i = 1, \ldots, n$ 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

$$X^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

Substitute into general formulae, and compute the estimate

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} =$$

$$= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}.$$

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,

$$\operatorname{var} b_1 = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

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

$$Y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}.$$

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*[4]. 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.

---

[4]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

Divisibility of natural numbers. 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:

$$a \mid b \wedge b \mid c \implies a \mid c,$$
$$a \mid b \wedge a \mid c \implies a \mid b + c \wedge a \mid b - c,$$
$$c \neq 0 \implies (a \mid b \Longleftrightarrow ac \mid bc),$$
$$a \mid b \wedge b > 0 \implies a \leq b.$$

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 $n - 1$.

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) (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.

We will introduce 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,[1]
- *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,
- *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: *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) iff there exists an integer c satisfying $a \cdot c = b$. We write this as $a \mid b$. The concept of divisibility can be defined (and its properties examined) much more generally – more can be found in 12.2.5.

---

[1]This question still belongs to open problems – 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.

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 $n = 2$ 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 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 into Bézout's identity can be determined this way (such integers k, l that $ka + lb = (a, b)$). It can also be easily proved straight from
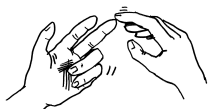
One of the most important properties of the integers, of which we will often take advantage, is the unique *Euclidean division* (division with remainder).

Theorem. *For any integers $a \in \mathbb{Z}$, $m \in \mathbb{N}$, there exists a unique pair of integers $q \in \mathbb{Z}$, $r \in \{0, 1, \ldots, m - 1\}$ 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 $a \in \mathbb{Z}$. First, we assume 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.

Now, suppose that $a \geq m$ and the existence of the integers q, r has been proved for all $a' \in \{0, 1, 2, \ldots, a - 1\}$. In particular, for $a' = a - m \geq 0$, there are q', r' such that $a' = q'm + r'$ and $r' \in \{0, 1, \ldots, m - 1\}$. 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 $q' \in \mathbb{Z}, r' \in \{0, 1, \ldots, m - 1\}$ 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 $a \in \mathbb{Z}, m \in \mathbb{N}$.

Now, we will prove the uniqueness. Suppose that there are integers $q_1, q_2 \in \mathbb{Z}$ and $r_1, r_2 \in \{0, 1, \ldots, m - 1\}$ which satisfy $a = q_1 m + r_1 = q_2 m + r_2$. 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

$$\frac{a}{m} = q + \frac{r}{m}, \quad \text{where} \quad 0 \leq \frac{r}{m} < 1.$$

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.

Consider integers a, b. An integer m satisfying both $m \mid a$ and $m \mid b$ is called a *common divisor* of a and b. A common divisor $m \geq 0$ of a and b which is divisible by every common divisor of the integers a, b is called the *greatest common divisor* of a and b and it is denoted by (a, b) (or $\gcd(a, b)$ for the sake of clarity).

The concept of the *least common multiple* is defined dually and denoted by $[a, b]$ (or $\operatorname{lcm}(a, b)$).

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:

$$10175 = 4 \cdot 2277 + 1067,$$
$$2277 = 2 \cdot 1067 + 143,$$
$$1067 = 7 \cdot 143 + 66,$$
$$143 = 2 \cdot 66 + 11,$$
$$66 = 6 \cdot 11 + 0.$$

Therefore, 11 is the greatest common divisor. We will express this integer from the particular equalities, resulting in a linear combination of the integers a, b:

$$11 = 143 - 2 \cdot 66$$
$$= 143 - 2 \cdot (1067 - 7 \cdot 143)$$
$$= -2 \cdot 1067 + 15 \cdot 143$$
$$= -2 \cdot 1067 + 15 \cdot (2277 - 2 \cdot 1067)$$
$$= 15 \cdot 2277 - 32 \cdot 1067$$
$$= 15 \cdot 2277 - 32 \cdot (10175 - 4 \cdot 2277)$$
$$= -32 \cdot 10175 + 143 \cdot 2277.$$

The wanted expression in the form of Bézout's identity is thus $11 = (-32) \cdot 10175 + 143 \cdot 2277$. □

The computation of the greatest common divisor using the Euclidean algorithm is quite fast even for relatively large integers. In our example, we will try this out with integers A, B, each of which will be the product of two 101-digit primes. Let us notice that the computation of the greatest common divisor of even such huge integers took an immeasurably small amount of time. A noticeable amount of time is taken by the computation of the greatest common divisor in the second example, where the input consists of two integers having more than a million digits.

An example in the system SAGE :

```
sage: p=next_prime(5*10^100)
sage: q=next_prime(3*10^100)
sage: r=next_prime(10^100)
sage: A=p*q;B=q*r;
sage: time G=gcd(A,B); print G
```

It follows straight from the definition that for any $a, b \in \mathbb{Z}$, we have $(a, b) = (b, a)$, $[a, b] = [b, a]$, $(a, 1) = 1$, $[a, 1] = |a|$, $(a, 0) = |a|$, $[a, 0] = 0$.

So far, we have not shown that for every pair of integers a, b, their greatest common divisor and least common multiple exist. However, if we assume they exist, then they are unique because every pair of non-negative integers k, l satisfy (directly from the definition) that $k \mid l$ and $l \mid k$ imply $k = l$. However, in the general case of divisibility in integral domains, the situation is more complicated – see 12.2.9. Even in the case of the so-called Euclidean domains,[2] which guarantee the existence of greatest common divisors, the result is determined uniquely up to the multiplication by a unit (an element having multiplicative inverse) – in the case of the integers, the result would be determined uniquely up to sign; the uniqueness was thus guaranteed by the condition that the greatest common divisor be non-negative.

Theorem (Euclidean algorithm). *Let a_1, a_2 be positive integers. For every $n \geq 3$ such that $a_{n-1} \neq 0$, let a_n denote the remainder of the division of a_{n-2} by a_{n-1}. Then, after a finite number of steps, we arrive at $a_k = 0$, and it holds that $a_{k-1} = (a_1, a_2)$.*

PROOF. By the Euclidean division, we have $a_2 > a_3 > a_4 > \ldots$. Since these are non-negative integers, this decreasing sequence cannot be infinite, so we get $a_k = 0$ after a finite number of steps, where $a_{k-1} \neq 0$. From the definition of the integers a_n, it follows that there are integers $q_1, q_2, \ldots, q_{k-2}$ such that

$$a_1 = q_1 \cdot a_2 + a_3,$$
$$a_2 = q_2 \cdot a_3 + a_4,$$
$$\vdots$$
$$a_{k-3} = q_{k-3} \cdot a_{k-2} + a_{k-1},$$
$$a_{k-2} = q_{k-2} \cdot a_{k-1}.$$

It follows from the last equality that $a_{k-1} \mid a_{k-2}$. Further, $a_{k-1} \mid a_{k-3}, \ldots, a_{k-1} \mid a_2, a_{k-1} \mid a_1$. Therefore, a_{k-1} is a common divisor of the integers a_1, a_2.

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 $a_4 = a_2 - q_2 a_3, a_5, \ldots$, 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. □

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.

Moreover, the Euclidean algorithm provides another interesting statement, which is often used.

11.1.3. Theorem (Bézout). *For every pair of integers a, b, there exist integers k, l such that $(a, b) = ka + lb$.*

_____

[2]Wikipedia, *Euclidean domain*, http://en.wikipedia.org/wiki/Euclidean_domain (as of July 29, 2017).

```
Time : CPU 0.00 s, Wall : 0.00 s
30000000000000000000000000000000000\
00000000000000000000000000000000000\
0000000000000000000000000000223

sage: time G=gcd(A^10000+1,B^10000+1);
Time : CPU 2.47 s, Wall : 2.48 s
```

11.A.5. Find the greatest common divisor of the integers $2^{49} - 1$ and $2^{35} - 1$, and determine the corresponding coefficients in Bézout's identity.

Solution. Again, we use the Euclidean algorithm. We get:

$$2^{49} - 1 = 2^{14}(2^{35} - 1) + 2^{14} - 1,$$
$$2^{35} - 1 = (2^{21} + 2^7)(2^{14} - 1) + 2^7 - 1,$$
$$2^{14} - 1 = (2^7 + 1)(2^7 - 1).$$

The wanted greatest common divisor is thus $2^7 - 1 = 127$. Let us notice that $7 = (49, 35)$ – see also the following exercise 11.A.6. Reversing this procedure, we find the coefficients k, ℓ into Bézout's identity $2^7 - 1 = k(2^{49} - 1) + \ell(2^{35} - 1)$:

$$2^7 - 1 = (2^{35} - 1) - (2^{21} + 2^7)(2^{14} - 1)$$
$$= (2^{35} - 1) - (2^{21} + 2^7)((2^{49} - 1) - 2^{14}(2^{35} - 1))$$
$$= (2^{35} + 2^{21} + 1)(2^{35} - 1) - (2^{21} + 2^7)(2^{49} - 1).$$

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.6. 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, ℓ satisfies $a^k - 1 \mid a^\ell - 1$ if and only if $k \mid \ell$. This can be proved by dividing the integer ℓ 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

$$a^{kq+r} - 1 = (a^k - 1)(a^{k(q-1)+r} + a^{k(q-2)+r} + \cdots + a^r) + a^r - 1$$

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. $\square$

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 $a_3 = a_1 - q_1 a_2$, $a_4 = a_2 - q_2 a_3$, ..., 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 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.A.7. 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. $\square$

Prime numbers. Euclid's theorem 11.2.1 is a fundamental property of primes as it can be seen in the following example.

11.A.8.

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.

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

$$(a_1, \ldots, a_n) = ((a_1, \ldots, a_{n-1}), a_n),$$
$$[a_1, \ldots, a_n] = [[a_1, \ldots, a_{n-1}], a_n].$$

Integers $a_1, a_2, \ldots, a_n \in \mathbb{Z}$ are said to be *coprime* (also relatively prime) iff $(a_1, a_2, \ldots, a_n) = 1$. They are said to be *pairwise coprime* (pairwise relatively prime) iff we have $(a_i, a_j) = 1$ for every pair of indices i, j satisfying $1 \le i < j \le n$.

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.

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. $\square$

11.A.9. 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.A.10. 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.

11.A.11. Prove that for any natural number n, there exist n consecutive natural numbers none of which is prime.

Solution. Let us examine the integers $(n + 1)! + 2$, $(n + 1)! + 3, \cdots, (n + 1)! + (n + 1)$. For any $k \in \{2, 3, \ldots, n+1\}$, 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. $\square$

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.

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)[3])

Remark. As mentioned in part 12.2.5, it is possible to define divisibility in a similar way in an arbitrary integral domain. In some integral domains (e.g. $\mathbb{Q}$), there are no elements with the prime property (we call them *irreducible*). 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 –

---

[3]Using the terminology of ring theory, it is a unit of the ring of the integers, which is usually assumed to be the product of the empty set of elements of the ring.

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[1] 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 $O\left(\log^3 N\right)$) 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.[2] 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).

11.A.12. Formulas for number and sum of divisors. 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

the elements which satisfy $p \mid ab \implies (p \mid a$ or $p \mid b)$ are always irreducible, but the contrary is not generally true. Let us mention at least one example of an ambiguous factorization – in $\mathbb{Z}[\sqrt{-5}]$, we have:[4] $6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5})$. However, it needs a longer discussion to verify that all of the mentioned factors are really irreducible in $\mathbb{Z}[\sqrt{-5}]$.

PROOF OF THE FUNDAMENTAL THEOREM OF ARITHMETIC. 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 $i \in \{1, \ldots, s\}, j \in \{1, \ldots, t\}$. We will prove by induction on s that $s = t$ and $p_1 = q_1, \ldots, p_s = q_s$.

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 $s - 1 = t - 1$, $p_1 = q_1, \ldots, p_{s-1} = q_{s-1}$. Altogether, we have $s = t$ and $p_1 = q_1, \ldots, p_{s-1} = q_{s-1}, p_s = q_s$. This proves the uniqueness, and thus the entire theorem as well. $\square$

11.2.3. Prime distribution.

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 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.

[1]M. Agrawal, N. Kayal, N. Saxena. *PRIMES is in P*. Annals of Mathematics 160 (2): 781–793. 2004.

[2]See http://www.rsasecurity.com/rsalabs/node.asp?id=2093.

[4]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.

factorizations. Every positive divisor of an integer $a = p_1^{\alpha_1} \cdot \cdots \cdot p_k^{\alpha_k}$ is of the form $p_1^{\beta_1} \cdot \cdots \cdot p_k^{\beta_k}$, where $\beta_1, \ldots, \beta_k \in \mathbb{N}_0$ and $\beta_1 \leq \alpha_1, \beta_2 \leq \alpha_2, \ldots, \beta_k \leq \alpha_k$.

Therefore, the integer a has exactly

$$\tau(a) = (\alpha_1 + 1)(\alpha_2 + 1) \cdot \cdots \cdot (\alpha_k + 1)$$

positive divisors, which sum up to

$$\sigma(a) = \frac{p_1^{\alpha_1+1} - 1}{p_1 - 1} \cdots \frac{p_k^{\alpha_k+1} - 1}{p_k - 1}.$$

Let $p_1, \ldots, p_k$ be pairwise distinct primes and $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$ be non-negative integers. Denoting $\gamma_i = \min\{\alpha_i, \beta_i\}$, $\delta_i = \max\{\alpha_i, \beta_i\}$ for every $i = 1, 2, \ldots, k$, then we have

$$(p_1^{\alpha_1} \cdot \cdots \cdot p_k^{\alpha_k}, p_1^{\beta_1} \cdot \cdots \cdot p_k^{\beta_k}) = p_1^{\gamma_1} \cdot \cdots \cdot p_k^{\gamma_k},$$
$$[p_1^{\alpha_1} \cdot \cdots \cdot p_k^{\alpha_k}, p_1^{\beta_1} \cdot \cdots \cdot p_k^{\beta_k}] = p_1^{\delta_1} \cdot \cdots \cdot p_k^{\delta_k}.$$

Solution. The justification of all of the mentioned propositions is a simple consequence of the first statement, which describes the factorizations of all of an integer's divisors. To find the number of positive divisors, we can use elementary combinatorics (the product rule) to get $\tau(a) = (\alpha_1+1)(\alpha_2+1) \cdot \cdots \cdot (\alpha_k+1)$. 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. $\square$

Perfect numbers and their relation to primes.

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 iff $\sigma(a) = 2a$, i.e., iff *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.A.13. 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.

Don Zagier

Now, we will try to answer 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 fundamental theorem, which we must not omit in this connection, is what Euclid around 300 BC was aware of:

11.2.4. 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 $p_1, p_2, \ldots, p_n$. Set $N = p_1 \cdot p_2 \ldots p_n + 1$. This integer, being greater than 1, is either a prime or it is divisible by a prime different from $p_1, \ldots, p_n$ (since the primes $p_1, \ldots, p_n$ divide the integer $N - 1$), which is a contradiction. $\square$

Now, 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[5].

Theorem (Chebyshev's, Bertrand's postulate). *For every integer $n > 1$, there is at least one prime p satisfying $n < p < 2n$.*

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.

11.2.5. 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 $1 \cdot m + a, 2 \cdot m + a, 3 \cdot m + a, \ldots$.*

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 $k \in \mathbb{N}_0$.*

Proof. Suppose the contrary, i.e., there are only finitely many primes of this form, and let them be denoted by $p_1 = 3$, $p_2 = 7$, $p_3 = 11$, $\ldots$, p_n. Further, let us set $N = 4p_2 \cdot p_3 \cdot \cdots p_n + 3$. 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

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

$$a = 2^k \cdot m, \text{ where } m, k \in \mathbb{N} \text{ and } 2 \nmid m.$$

Since the function σ 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. $\qquad\square$

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

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 $p_1, p_2, \ldots, p_n$ can play the role of the mentioned p since if we had $p_i \mid N$ for some $i \in \{2, \ldots, n\}$, 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. $\qquad\square$

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).

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.

11.2.6. 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 $x \to \infty$.

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.

Proposition. *Let P denote the set of all primes, then*

$$\sum_{p \in P} \frac{1}{p} = \infty.$$

Remark. On the other hand, $\sum_{n \in \mathbb{N}} \frac{1}{n^2} = \frac{\pi^2}{6}$, which means that the primes are distributed more "densely" in $\mathbb{N}$ than squares.

Proof. Let n be an arbitrary natural number and $p_1, \ldots, p_{\pi(n)}$ all primes less than or equal to n. Let us set

$$\lambda(n) = \prod_{i=1}^{\pi(n)} \left(1 - \frac{1}{p_i}\right)^{-1}.$$

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.)

B. 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 proofs will often be much shorter and clearer.

11.B.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) $m \mid a - b$.

Solution. (1) $\Longrightarrow$ (3) If $a = q_1 m + r$, $b = q_2 m + r$, then $a - b = (q_1 - q_2)m$.

(3) $\Longrightarrow$ (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) $\Longrightarrow$ (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.B.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

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \cdots + b^{n-1}),$$

it follows that $a^n \equiv b^n \pmod{m}$ as well.

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 $(\alpha_1, \ldots, \alpha_{\pi(n)})$. Since every integer not exceeding n factors to only primes in the set $\{p_1, \ldots, p_{\pi(n)}\}$, all of them are included in this sum. Therefore, $\lambda(n) > 1 + \frac{1}{2} + \cdots + \frac{1}{n}$, and since the harmonic series is divergent (see 5.H.3), we also have $\lim_{n \to \infty} \lambda(n) = \infty$.

Taking into account the expansion of the function $\ln(1 + x)$ to a power series (see 6.C.1), we further get

$$\ln \lambda(n) = -\sum_{i=1}^{\pi(n)} \ln \left(1 - \frac{1}{p_i} \right) = \sum_{i=1}^{\pi(n)} \sum_{m=1}^{\infty} (m p_i^m)^{-1} =$$

$$= p_1^{-1} + \cdots + p_{\pi(n)}^{-1} + \sum_{i=1}^{\pi(n)} \sum_{m=2}^{\infty} (m p_i^m)^{-1}.$$

Since the inner sum can be bound from above by

$$\sum_{m=2}^{\infty} (m p_i^m)^{-1} < \sum_{m=2}^{\infty} p_i^{-m} =$$

$$= p_i^{-2} \left(1 - p_i^{-1} \right)^{-1} \leq 2 p_i^{-2},$$

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

$$a \not\equiv b \pmod{m}.$$

Whenever it is apparent that we are working with congruence relations, we usually omit the symbol mod and write just $a \equiv b \, (m)$.

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 $a \equiv b \pmod{m}$, $b = b_1 d$, $m = m_1 d$. Then there is an integer t such that $a = b + mt = b_1 d + m_1 dt = (b_1 + m_1 t)d$, hence $d \mid a$.

vii) If $a \equiv b \pmod{m_1}, a \equiv b \pmod{m_2}, \ldots, a \equiv b \pmod{m_k}$, then the difference $a - b$ is a common multiple of the integers $m_1, m_2, \ldots, m_k$, and so it is divisible by their least common multiple $[m_1, m_2, \ldots, m_k]$, whence it follows that $a \equiv b \pmod{[m_1, \ldots, m_k]}$. $\square$

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 convenient way of writing which simplifies both expressions and reasonings.

11.B.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

11.3.1. Fundamental properties of congruences. It follows directly from the definition that the congruence modulo m is an *equivalence* relation.

Now, we will state further properties of congruences which will be proven in the practical column.

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 $m_1, \ldots, m_k$, it is also valid with respect to their least common multiple $[m_1, \ldots, m_k]$.

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) of 11.3.1.

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. Möbius inversion formula and Euler function.

Here, an arithmetic function means any function whose domain is the set of natural numbers.

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).

Example. $\mu(4) = \mu(2^2) = 0, \mu(6) = \mu(2\cdot3) = (-1)^2 = 1$, $\mu(2) = \mu(3) = -1$.

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). □

We now show how helpful the notion of congruence can be in solving problems similar to one already solved in 11.A.7 (where the induction and binomial theorem were used).

11.B.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,

$$37^{n+2} + 16^{n+1} + 23^n \equiv 2^{n+2} + 2^{n+1} + 2^n$$
$$= 2^n(4+2+1) \equiv 0 \pmod 7. \qquad \square$$

11.B.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

$$n \equiv (2^5 + 6)^{18} - 1 = 38^{18} - 1 \equiv 3^{18} - 1$$
$$= 27^6 - 1 \equiv (-1)^6 - 1 = 0 \pmod 7,$$

hence $7 \mid n$. Similarly, $835 \equiv 3 \pmod{16}$, so

$$n \equiv (3^5 + 6)^{18} - 1 = (3 \cdot 81 + 6)^{18} - 1 \equiv (3 \cdot 1 + 6)^{18} - 1$$
$$= 9^{18} - 1 = 81^9 - 1 \equiv 1^9 - 1 = 0 \pmod{16},$$

hence $16 \mid n$. Altogether, $112 \mid n$, which was to be proved. □

11.B.6. Prove that the following relations hold for prime p:

i) If $k \in \{1, \ldots, p-1\}$, 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.

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 $p \mid \binom{p}{k}$.

Now, we prove several important properties of the Möbius function, especially the so-called *Möbius inversion formula*.

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 $i \in \{1, \ldots, k\}$. Therefore,

$$\sum_{d \mid n} \mu(d) = \sum_{\substack{(\beta_1, \ldots, \beta_k) \in (\mathbb{N} \cup \{0\})^k \\ 0 \leq \beta_i \leq \alpha_i}} \mu\left(p_1^{\beta_1} \cdots p_k^{\beta_k}\right)$$

$$= \sum_{(\beta_1, \ldots, \beta_k) \in \{0,1\}^k} \mu\left(p_1^{\beta_1} \cdots p_k^{\beta_k}\right)$$

$$= \binom{k}{0} + \binom{k}{1} \cdot (-1) + \binom{k}{2} \cdot (-1)^2 + \cdots + \binom{k}{k} \cdot (-1)^k$$

$$= (1 + (-1))^k = 0.$$

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 ℓ of the exponents $\beta_1, \ldots, \beta_k$ are equal to one; there are $\binom{k}{\ell}$ of them, and each satisfies that $\mu(p_1^{\beta_1} \cdots p_k^{\beta_k}) = (-1)^\ell$. □

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:

$$(f \circ g)(n) = \sum_{d \mid n} f(d) \cdot g\left(\tfrac{n}{d}\right) = \sum_{d_1 d_2 = n} f(d_1) \cdot g(d_2).$$

Lemma. *The Dirichlet product is associative.*

PROOF.

$$((f \circ g) \circ h)(n) = \sum_{d_1 d_2 d_3 = n} f(d_1)g(d_2)h(d_3) = (f \circ (g \circ h))(n)$$

□

Example. Let us define two helping 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:

$$f \circ i = i \circ f = f \quad \text{and} \quad (I \circ f)(n) = (f \circ I)(n) = \sum_{d \mid n} f(d).$$

Further, $I \circ \mu = \mu \circ I = i$, since

$$(I \circ \mu)(n) = \sum_{d \mid n} I(d) \mu\left(\tfrac{n}{d}\right) = \sum_{d \mid n} I\left(\tfrac{n}{d}\right) \mu(d) =$$

$$= \sum_{d \mid n} \mu(d) = 0 \quad \text{for all } n > 1$$

by the lemma after the definition of the Möbius function (the statement is clearly true for $n = 1$).

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 $k \in \{1, \ldots, p-1\}$, whence the statement follows easily. $\square$

11.B.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 $a^{m-1} + a^{m-2} + \cdots + b^{m-1}$ 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.B.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 $a^4 \equiv 1^2 \pmod{2^4}$. $\square$

11.B.9. Divisibility rules. We can surely recall the basic rules of divisibility (at least by the numbers 2, 3, 4, 5, 6, 9 a 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?

Theorem (Möbius inversion formula)**.** *Let an arithmetic function F be defined in terms of an arithmetic function f by $F(n) = \sum_{d|n} f(d)$. Then, the function f can be expressed as*

$$f(n) = \sum_{d|n} \mu\left(\tfrac{n}{d}\right) \cdot F(d).$$

Proof. The relation $F(n) = \sum_{d|n} f(d)$ 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. $\square$

Definition. A *multiplicative function* on the natural numbers is such an arithmetic function which, for all pairs of *coprime* natural numbers a, b, satisfies

$$f(a \cdot b) = f(a) \cdot f(b).$$

Example. Multiplicative functions include, for instance, $\sigma(n)$, $\tau(n)$, $\mu(n)$ (this can be verified easily from the definition) or, as we will prove shortly, the so-called (Euler) totient function $\varphi(n)$.

Remark. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci. Tady by mela byt poznamka (by T. Perutka) o Okruhu aritmetickych funkci. Zaroven trochu vyplni prostor, ktery je potreba kvuli velkemu mnozstvi prikladu na Eulerovu funkci v praktickem sloupci.

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).

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 $n = a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_1 10 + a_0$ is equal to $S(n) = a_k + a_{k-1} + \cdots + a_0$, and since $10^\ell \equiv 1^\ell = 1 \pmod 9$ for any $\ell \in \mathbb{N}_0$, we get

$$n = a_k 10^k + \cdots + a_0 \equiv a_k + \cdots + a_0 = S(n) \pmod 9.$$

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

$$n = a_k 10^k + \cdots + a_0 \equiv a_k(-1)^k + \cdots + a_1(-1) + a_0$$
$$\equiv (a_0 + a_2 + \cdots) - (a_1 + a_3 + \cdots) \pmod{11}.$$

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 12.4.1 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

For a natural number, we define the value of Euler's totient function φ as

$$\varphi(n) = |\{a \in \mathbb{N} \mid 0 < a \leq n, \, (a, n) = 1\}|$$

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).

Now, we are going to prove several important properties of the φ function.

Lemma. Let $n \in \mathbb{N}$. Then, $\sum_{d|n} \varphi(d) = n$.

PROOF. Let us consider the n fractions

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}.$$

Reducing them to lowest terms and grouping them together by the denominators, we get just the statement in question. $\square$
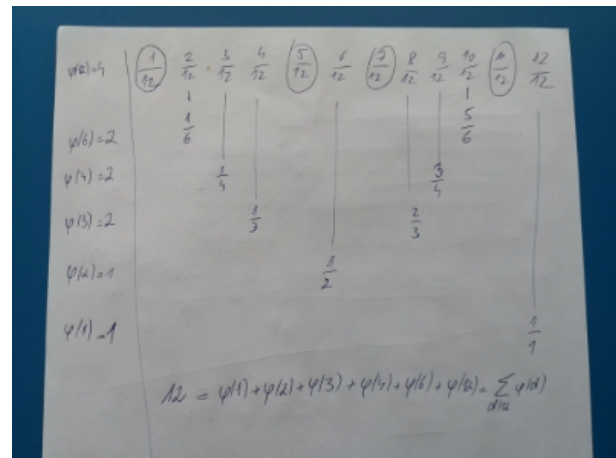
Theorem. Let $n \in \mathbb{N}$ factor to primes as $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$. Then,

$$\varphi(n) = n \cdot \left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_k}\right).$$

PROOF. Invoking the previous lemma and the Möbius inversion formula, we get

$$\varphi(n) = \sum_{d|n} \mu(d) \frac{n}{d}$$
$$= n - \frac{n}{p_1} - \cdots - \frac{n}{p_k} + \cdots + (-1)^k \frac{n}{p_1 \cdots p_k}$$
$$= n \cdot \left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_k}\right).$$

$\square$

Remark. The previous result can also be obtained without using the Möbius inversion formula by the inclusion-exclusion principle, determining the number of integers not coprime to n in a given interval.

It follows directly from this theorem that the totient function is a multiplicative arithmetic function.

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):

$$0 \equiv b_9 2^9 + b_8 2^8 + b_7 2^7 + \cdots + b_3 2^3 + b_2 2^2 + b_1 2^1 + b_0 2^0$$
$$\equiv -5b_9 + 3b_8 - 4b_7 - 2b_6 - b_5$$
$$+ 5b_4 - 3b_3 + 4b_2 + 2b_1 + b_0 \quad (\mathrm{mod}\ 11).$$

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.

11.B.10. Verify that the account number of the Masaryk university, 85636621, is built correctly.

Solution. We will test the condition of divisibility by eleven:
$$-5b_9 + 3b_8 - 4b_7 - 2b_6 - b_5 + 5b_4 - 3b_3 + 4b_2 + 2b_1 + b_0$$
$$\equiv -4 \cdot 8 - 2 \cdot 5 - 1 \cdot 6 + 5 \cdot 3 - 3 \cdot 6 + 4 \cdot 6 + 2 \cdot 2 + 1 \cdot 1$$
$$\equiv \quad 0\ (\mathrm{mod}\ 11)\,. \qquad \square$$

Euler's totient function. The totient function φ 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 $\varphi(m) = |\{a \in \mathbb{N} \mid 0 < a \le m,\ (a, m) = 1\}|$.

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.B.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 $\varphi(72) = \varphi(8) \cdot \varphi(9) = 4 \cdot 6 = 24$. $\qquad \square$

11.B.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 $n \ge 3$ 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

Corollary. *Let* $a, b \in \mathbb{N}$, $(a, b) = 1$. *Then*
$$\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b).$$

Remark. This statement can also be derived independently of the knowledge that $(n, ab) = 1 \iff (n, a) = 1 \wedge (n, b) = 1$. Together with the easy result
$$\varphi(p^\alpha) = p^\alpha - p^{\alpha - 1} = (p - 1) \cdot p^{\alpha - 1},$$
one can deduce the formula for the computation of φ in a third way.

Möbius inversion formula and irreducible polynomials. Above, we prove the properties of Euler's totient function using Möbius inversion formula.

The standard form of this formula connects the expression of an arithmetic function F of natural numbers in terms of a function f in the form
$$F(n) = \sum_{d \mid n} f(d)$$
to the inverse expression of the function f in terms of the function F in the form
$$f(n) = \sum_{d \mid n} \mu(\tfrac{n}{d}) \cdot F(d).$$

The function value $\mu(n)$ depends on the prime factorization of the input value n as follows:

- if a square of a prime divides n, then $\mu(n) = 0$;
- otherwise, we set $\mu(n) = (-1)^k$, where k is the number of primes which divide n.

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
$$f(n) = \sum_{d \mid n} F(d)^{\mu\left(\frac{n}{d}\right)}.$$

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
$$x^{p^n} - x = \prod_{d \mid n} S_d(x).$$

Confronting the degrees of the polynomial on both sides yields
$$p^n = \sum_{d \mid n} d N_d,$$

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 $\varphi(4n + 2) = \varphi(2 \cdot (2n + 1)) = \varphi(2) \cdot \varphi(2n + 1) = \varphi(2n + 1)$. □

11.B.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 $p^{\alpha - 1} \mid a$.

- 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 $m \in \{31, 62\}$.

whence we get, by applying the standard Möbius inversion formula, that

$$N_n = \frac{1}{n} \sum_{d \mid n} \mu\left(\frac{n}{d}\right) p^d.$$

In particular, we can see that for any $n \in \mathbb{N}$, it holds that $N_n = \frac{1}{n}\left(p^n - \cdots + \mu(n)p\right) \neq 0$ since the expression in the parentheses is a sum of distinct powers of p multiplied by coefficients ± 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).

11.3.3. Example. By the formula we have proved, the number of (monic) irreducible polynomials over $\mathbb{F}_2$ of degree 5 is equal to

$$N_5 = \frac{1}{5} \sum_{d \mid 5} \mu\left(\tfrac{5}{d}\right) 2^d = \frac{1}{5}\left(\mu(1) \cdot 2^5 + \mu(5) \cdot 2\right) = 6.$$

The number of monic irreducible polynomials over $\mathbb{F}_3$ of degree four is then

$$N_4 = \frac{1}{4} \sum_{d \mid 4} \mu\left(\tfrac{4}{d}\right) 3^d = \frac{1}{4}\left(\mu(1) \cdot 3^4 + \mu(2) \cdot 3^2 + \mu(4)3^1\right)$$
$$= \frac{1}{4}(81 - 9) = 18 .$$

11.3.4. Fermat's little theorem, Euler's theorem. These

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.

Theorem (Fermat's little). *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 12.3.10). However, it can be proved directly (by mathematical induction or a combinatorial means, as mentioned in exercise 11.B.15). □

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,*

$$a^p \equiv a \pmod{p}.$$

Before formulating and proving Euler's theorem, we introduce a few useful concepts.

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, $\varphi(m) = 2 \cdot 3^{\alpha-1} \cdot \varphi(n) = \frac{m}{3} = 3^{\alpha-1} \cdot n$. Reducing this leads to $2\varphi(n) = n$ or, equivalently, $\varphi(n) = \frac{n}{2}$. Here, we must have $2 \mid n$, and writing $n = 2^\beta \cdot k$, where $(k, 6) = 1, \beta \leq 1$, we get $\varphi(k) = k$, which is apparently satisfied only by $k = 1$.

To summarize it, the problem is satisfied by those natural numbers which are of the form $2^\alpha 3^\beta$, where $\alpha, \beta \geq 1$. $\square$

11.B.14. Find all two-digit numbers n for which $9 \mid \varphi(n)$. ○

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 $0, 1, \ldots, m - 1$ or, for odd m, its "symmetric" variation $-\frac{m-1}{2}, \ldots, -1, 0, 1, \ldots, \frac{m-1}{2}$).

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 $x_1, x_2, \ldots, x_{\varphi(m)}$ form a reduced residue system modulo m. If $a \in \mathbb{Z}, (a, m) = 1$, then the integers $a \cdot x_1, \ldots, a \cdot x_{\varphi(m)}$ 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. $\square$

Theorem (Euler). *Let $a \in \mathbb{Z}, m \in \mathbb{N}, (a, m) = 1$. Then,*

$$a^{\varphi(m)} \equiv 1 \pmod{m}.$$

PROOF. Let $x_1, x_2, \ldots, x_{\varphi(m)}$ be an arbitrary reduced residue system modulo m. By the previous lemma, $a \cdot x_1, \ldots, a \cdot x_{\varphi(m)}$ is also a reduced residue system modulo m. Therefore, for every $i \in \{1, 2, \ldots, \varphi(m)\}$, there is a unique $j \in \{1, 2, \ldots, \varphi(m)\}$ 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) 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 $(\mathbb{Z}_m^\times, \cdot)$.

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:

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

11.B.15. Fermat's (little) theorem. Now, we will prove Fermat's little theorem 11.3.4 in two more ways: by mathematical induction, and then by a combinatorial means). The theorem states that for any integer a and any prime p which does not divide a, it holds that $a^{p-1} \equiv 1 \pmod{p}$.

Solution. First, we prove (by induction on a) that an apparently equivalent statement, $a^p \equiv a \pmod{p}$, holds for any $a \in \mathbb{Z}$ and prime p. For $a = 1$, there is nothing to prove. Further, let us assume that the proposition holds for a and prove its validity for $a + 1$. It follows from the induction hypothesis and the exercise 11.B.6 that

$$(a + 1)^p \equiv a^p + 1^p \equiv a + 1 \pmod{p},$$

which is what we were to prove.

The statement holds trivially for $a = 0$ as well as in the case $a < 0, p = 2$. The validity for $a < 0$ and p odd can be obtained easily from the above: since $-a$ is a positive integer, we get $-a^p = (-a)^p \equiv -a \pmod{p}$, whence $a^p \equiv a \pmod{p}$.

The combinatorial proof is a somewhat "cunning" one: Similarly to problems using Burnside's lemma (see exercise 12.G.1), we are to determine how many necklaces can be created by wiring a given number of beads, of which there is a given number of types. Having a types of beads, there are clearly a^p necklaces of length p, a of which consist of a single bead type. From now on, we will be interested only in the other ones, of which there are thus $a^p - a$. Apparently, each necklace is transformed into itself by rotating by p beads. In general, a necklace can be transformed into itself by rotating by another number of beads, but this number can never be coprime to p (for instance, considering $p = 8$ and the necklace $ABABABAB$, rotations by 2, 4, or 6 beads leave it unchanged). However, if p is a prime, it follows that all rotations lead to different necklaces. Therefore, if we do not distinguish necklaces which differ in rotation only (i.e., in the position of the "knot"), there are exactly

$$\frac{a^p - a}{p}$$

of them, which especially means that $p \mid a^p - a$.

As an example, let us consider the case $a = 2, p = 5$, i.e., necklaces of length 5, consisting of 2 bead types (A, B). There are $2^5 = 32$ necklaces in total, 2 of which consist of a single bead type $(AAAAA, BBBBB)$. Leaving them and ignoring the position of the knot, there remain $\frac{2^5 - 2}{5} = 6$

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 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.B.19, and complete description of the dependency of the order upon the exponent is given by the subsequent two theorems.

11.3.5. 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 $q, z \in \mathbb{N}_0, 0 \leq z < r$.

"$\Longleftarrow$" 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.

"$\Longrightarrow$" 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 $z < r$, 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 12.3.10 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) $r \mid \varphi(m)$

The following theorem is a generalization of the result in 11.B.20.

11.3.6. 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 $\frac{r}{(n, r)}$.*

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. $\square$

necklaces which differ not merely in rotation, namely *ABBBB, AABBB, AAABB, AAAAB, ABABB, AABAB.* □

11.B.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 $2^{16} \equiv 3^{16} \equiv 4^{16} \equiv 1$ (mod 17). Since $50 \equiv 2$ (mod 16), we get
$$2^{50} + 3^{50} + 4^{50} \equiv 2^2 + 3^2 + 4^2 \equiv 12 \pmod{17}.$$

ii) Similarly $2^{36} \equiv 3^{36} \equiv 5^{36} \equiv 1$ (mod 37), and hence
$$2^{181} + 3^{181} + 5^{181} \equiv 2 + 3 + 5 \equiv 10 \pmod{37}. \quad □$$

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.B.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.B.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.B.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 $a^n \equiv 1 \pmod{m} \iff b^n \equiv 1 \pmod{m}$. □

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 δ 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, $\delta = rs$. □

11.3.7. 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).

PRIMITIVE ROOT

Let $m \in \mathbb{N}$. An integer g, $(g, m) = 1$, is said to be a *primitive root* modulo m iff 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 $x_a \in \mathbb{Z}$, $0 \le x_a < \varphi(m)$ 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 $\{x \in \mathbb{Z}; 0 \le x < \varphi(m)\}$.

PROOF. Suppose that it holds for $x, y \in \mathbb{Z}$, $0 \le x, y < \varphi(m)$ 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. □

If there are primitive roots at all for a natural number m, then there are exactly $\varphi(\varphi(m))$ of them among the integers $1, 2, \ldots, m$: If g is a primitive root and $a \in \{1, 2, \ldots, \varphi(m)\}$ arbitrary, then the order of g^a is $\frac{\varphi(m)}{(a, \varphi(m))}$ (by theorem 11.3.6), 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 $\{1, 2, \ldots, \varphi(m)\}$.

Now, we are about to show that primitive roots exist for a sufficient amount of moduli m.

11.B.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 $a, a^2, a^3, \ldots, a^{rs-1}$ is congruent to 1 modulo m, nor is any of the integers $a^r, a^{2r}, a^{3r}, \ldots, a^{(s-1)r}$. On the other hand, we have $(a^r)^s \equiv 1 \pmod{m}$, so the order of a^r modulo m equals s. $\square$

11.B.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.B.22. Determine the last digit of the numbers

i) $3^{5^{7^9}}$,

ii) $37^{37^{37}}$,

iii) $12^{13^{14}}$. $\bigcirc$

11.B.23. It holds for all odd $n \in \mathbb{N}$ that $n \mid 2^{n!} - 1$. Prove this! $\bigcirc$

11.B.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.

Solution.

i) The order of 7 modulo 100 is equal to 4 by exercise 11.B.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.B.28); then, theorem 11.3.6 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 finish the calculation as follows: $4^4 \equiv 4^{-1} \equiv 3 \pmod{11}$.

11.3.8. 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 12.3.8).

Proposition. *Let p be an odd prime. Then there are primitive roots modulo p.*

Proof. Let $r_1, r_2, \ldots, r_{p-1}$ be the orders of the integers $1, 2, \ldots, p-1$ modulo p. Let $\delta = [r_1, r_2, \ldots, r_{p-1}]$ be the least common multiple of these orders. We will show that there is an integer of order δ among $1, 2, \ldots, p-1$ and that $\delta = p - 1$.

Let $\delta = q_1^{\alpha_1} \cdots q_k^{\alpha_k}$ be the factorization of δ to primes. For every $s \in \{1, \ldots, k\}$, there is a $c \in \{1, \ldots, p-1\}$ such that $q_s^{\alpha_s} \mid r_c$ (otherwise, there would be a common multiple of the integers $r_1, r_2, \ldots, r_{p-1}$ less than δ). 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 11.3.6 on orders of powers).

Reasoning analogously for any $s \in \{1, \ldots, k\}$, we get integers $g_1, \ldots, g_k$, 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 $g_1, \ldots, g_k$, 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 $1, 2, \ldots, p-1$ divide δ, we get the congruence $x^\delta \equiv 1 \pmod{p}$ for any $x \in \{1, 2, \ldots, p-1\}$. By theorem 11.4.8, there are at most δ solutions to a congruence of degree δ 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 12.2.4). On the other hand, we have already shown that this congruence has $p - 1$ solutions, so necessarily $\delta \geq p - 1$. Still, δ is (being the order of g) a divisor of $p - 1$, whence we finally get the wanted equality $\delta = p - 1$. $\square$

Now, we show that there are primitive roots modulo powers of odd primes. First, we prove two helping lemmas.

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 ℓ.

I. The statement is clearly true for $\ell = 2$.

II. Let the statement be true for ℓ, and let us prove it for $\ell + 1$.

11.B.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,

$$14^{14^{14}} \equiv 14^{16} = 2^{16} \cdot 7^{16} \pmod{25}.$$

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,

$$14^{14^{14}} \equiv 2^{16} \cdot 7^{16} \equiv (2^5)^3 \cdot 2 \cdot 7^{16} \equiv 7^3 \cdot 2 \cdot 7^{16}$$
$$\equiv 2 \cdot 7^{19} \equiv 2 \cdot (-1)^9 \cdot 7 = 11 \pmod{25}.$$

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.B.26. Determine the last three digits of the number $12^{10^{11}}$.

11.B.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	$\mathrm{mod}\ 11$	1	5	3	4	9
4^n	$\mathrm{mod}\ 11$	1	4	5	9	3
3^n	$\mathrm{mod}\ 11$	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 $n \equiv 2 \pmod 5$. $\square$

Invoking exercise 11.B.7 and raising the statement for ℓ 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.B.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. $\square$

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^ℓ 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^ℓ, 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 $(1 + ap)^{p^{\ell-2}} \not\equiv 1 \pmod{p^\ell}$, 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^ℓ.*

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^ℓ 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^ℓ 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^ℓ is $p^{\ell-1}$. However, then it follows from the corollary of 11.3.5 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^ℓ. $\square$

Proposition. *Let p be an odd prime and g a primitive root modulo p^ℓ 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^ℓ is also a primitive root modulo $2p^\ell$. $\square$

11.B.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. □

11.B.29. We will now determine (with the help of propositions proving Theorem 11.3.8) 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

$$g^{20} \not\equiv 1 \pmod{41} \land g^8 \not\equiv 1 \pmod{41}.$$

Now, we will go through the potential primitive roots in ascending order:

$$
\begin{aligned}
g = 2: \quad & 2^8 = 2^5 \cdot 2^3 \equiv -9 \cdot 8 \equiv 10 \pmod{41}, \\
& 2^{20} = (2^5)^4 \equiv (-9)^4 = 81^2 \equiv (-1)^2 \\
& \quad = 1 \pmod{41}, \\
g = 3: \quad & 3^8 = (3^4)^2 \equiv (-1)^2 = 1 \pmod{41}, \\
g = 4: \quad & \text{the order of } 4 = 2^2 \text{ always divides the order 2,} \\
g = 5: \quad & 5^8 = (5^2)^4 \equiv (-2^4)^4 = 2^{16} = (2^8)^2 \\
& \quad \equiv 10^2 \equiv 18 \pmod{41}, \\
& 5^{20} = (5^2)^{10} \equiv (-2^4)^{10} = 2^{40} = (2^{20})^2 \\
& \quad \equiv 1 \pmod{41}, \\
g = 6: \quad & 6^8 = 2^8 \cdot 3^8 \equiv 10 \cdot 1 = 10 \pmod{41}, \\
& 6^{20} = 2^{20} \cdot 3^{20} \equiv 2^{20} \cdot (3^8)^2 \cdot 3^4 \\
& \quad \equiv 1 \cdot 1 \cdot (-1) = -1 \pmod{41}.
\end{aligned}
$$

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

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 $2 \nmid p$. □

Lemma. *Let $\ell \in \mathbb{N}, \ell \geq 3$. Then the order of the integer 5 modulo 2^ℓ is $2^{\ell-2}$.*

PROOF. Easily from the above lemma. □

Proposition. *Let $\ell \in \mathbb{N}$. There are primitive roots modulo 2^ℓ 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^ℓ: it has $\varphi(2^\ell)$ elements, and it can be easily verified that they are pairwise incongruent modulo 2^ℓ.

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 $\varphi(2^\ell) = 2^{\ell-1}$. □

The last piece to the jigsaw puzzle of propositions which collectively prove theorem 11.3.8 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 $\delta = [\varphi(2^\alpha), \varphi(p_1^{\alpha_1}), \ldots, \varphi(p_1^{\alpha_1})]$, 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. □

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.

11.3.9. 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 $q_1, \ldots, q_k$ are primes and $\alpha_1, \ldots, \alpha_k \in \mathbb{N}$. 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:*

$$g^{\frac{\varphi(m)}{q_1}} \equiv 1 \pmod m, \ldots, g^{\frac{\varphi(m)}{q_k}} \equiv 1 \pmod m.$$

PROOF. If either of the congruences were true, it would mean that the order of g is less than $\varphi(m)$.

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:

$$2^8 = 256 = 6 \cdot 41 + 10 \quad (\bmod\ 41^2),$$
$$3^8 = (3^4)^2 = (2 \cdot 41 - 1)^2 \equiv -4 \cdot 41 + 1 \quad (\bmod\ 41^2).$$

Then,

$$6^8 = 2^8 \cdot 3^8 \equiv (6 \cdot 41 + 10)(-4 \cdot 41 + 1)$$
$$\equiv -34 \cdot 41 + 10 \equiv 7 \cdot 41 + 10 \quad (\bmod\ 41^2)$$

and

$$6^{40} = (6^8)^5 \equiv (7 \cdot 41 + 10)^5 \equiv (10^5 + 5 \cdot 7 \cdot 41 \cdot 10^4)$$
$$= 10^4(10 + 35 \cdot 41) \equiv (-2 \cdot 41 - 4)(-6 \cdot 41 + 10)$$
$$\equiv (4 \cdot 41 - 40) = 124 \not\equiv 1 \quad (\bmod\ 41^2).$$

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). $\square$

C. 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.
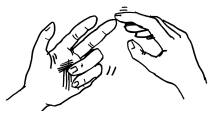
11.C.1. Solve the congruence

$$39x \equiv 41 \quad (\bmod\ 47).$$

Solution.

i) First, we use Euler's theorem.
Since $(39, 47) = 1$, we have
$$39^{\varphi(47)} = 39^{46} \equiv 1 \quad (\bmod\ 47),$$
i.e.,
$$\underbrace{39^{45} \cdot 39}_{39^{46} \equiv 1} x \equiv 39^{45} \cdot 41 \quad (\bmod\ 47),$$

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 $i \in \{1, \ldots, k\}$ such that $q_i \mid u$. However, then we get

$$g^{\frac{\varphi(m)}{q_i}} = g^{d \cdot \frac{u}{q_i}} \equiv 1 \quad (\bmod\ m).$$
$\square$

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.

The mentioned congruence is equivalent to the congruence

$$f(x) - g(x) \equiv 0 \quad (\bmod\ m).$$

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*

$$a \equiv b \quad (\bmod\ m) \implies f(a) \equiv f(b) \quad (\bmod\ m).$$

PROOF. Let $f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + + c_0$, where $c_0, c_1, \ldots, c_n \in \mathbb{Z}$. Since $a \equiv b \pmod{m}$, $c_i a^i \equiv c_i b^i \pmod{m}$ holds for every $i = 0, 1, \ldots, n$. Adding up these congruences for $i = 0, 1, 2, \ldots, n$ leads to

$$c_n a^n + \cdots + c_1 a + c_0 \equiv c_n b^n + \cdots + c_1 b + c_0 \quad (\bmod\ m),$$

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.

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

$$47 = 1 \cdot 39 + 8,$$
$$39 = 4 \cdot 8 + 7,$$
$$8 = 1 \cdot 7 + 1.$$

In the other direction, this leads to

$$1 = 8 - 7 = 8 - (39 - 4 \cdot 8) = 5 \cdot 8 - 39$$
$$= 5 \cdot (47 - 39) - 39 = 5 \cdot 47 - 6 \cdot 39.$$

Considering this equality modulo 47 and remembering that we are solving the equation $41 \equiv x \cdot 39$, we obtain

$$1 \equiv -6 \cdot 39 \qquad \pmod{47}, \qquad / \cdot 41$$
$$41 \equiv \underbrace{41 \cdot (-6)} \cdot 39 \qquad \pmod{47},$$
$$x \equiv 41 \cdot (-6) \qquad \pmod{47},$$
$$x \equiv -246 \qquad \pmod{47},$$
$$x \equiv 36 \qquad \pmod{47}.$$

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.

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:

$$39x \equiv 41 \qquad \pmod{47},$$
$$-8x \equiv -6 \qquad \pmod{47}, \qquad / : -2$$
$$4x \equiv 3 \qquad \pmod{47},$$
$$4x \equiv -44 \qquad \pmod{47}, \qquad / : 4$$
$$x \equiv -11 \qquad \pmod{47},$$
$$x \equiv 36 \qquad \pmod{47}.$$

□

(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)*

$$ax \equiv b \pmod m$$

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. □

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

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.C.2. Solve the system

$$x \equiv 7 \pmod{27},$$
$$x \equiv -3 \pmod{11}.$$

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.C.3. Solve the following system of congruences:

$$x \equiv 1 \pmod{10},$$
$$x \equiv 5 \pmod{18},$$
$$x \equiv -4 \pmod{25}.$$

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):

$$1 + 10t \equiv 5 \pmod{18},$$
$$10t \equiv 4 \pmod{18},$$
$$5t \equiv 2 \pmod{9},$$
$$5t \equiv 20 \pmod{9},$$
$$t \equiv 4 \pmod{9},$$

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 $x = 1 + 10t = 1 + 10(4 + 9s) = 41 + 90s$.

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 $\{2, 3, \ldots, p - 2\}$ 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

$$(p - 1)! \equiv 1^{(p-3)/2} \cdot (p - 1) \equiv -1 \pmod p. \quad □$$

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

$$x \equiv c_1 \pmod{m_1},$$
$$\vdots$$
$$x \equiv c_k \pmod{m_k}.$$

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*

$$x \equiv c_1 \pmod{m_1},$$
$$x \equiv c_2 \pmod{m_2}$$

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.,

Once again, this can be substituted into the third congruence and then solved:

$$41 + 90s \equiv -4 \pmod{25},$$
$$90s \equiv 5 \pmod{25},$$
$$18s \equiv 1 \pmod{5},$$
$$3s \equiv 6 \pmod{5},$$
$$s \equiv 2 \pmod{5},$$

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 $x \equiv 221 \pmod{450}$. $\qquad\square$

11.C.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:

$$x \equiv 10 \pmod{13},$$
$$x \equiv 3 \pmod{12},$$
$$x \equiv 0 \pmod{11}.$$

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. $\qquad\square$

11.C.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?

$tm_1 \equiv c_2 - c_1 \pmod{m_2}$. 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

$$t \equiv \frac{c_2 - c_1}{d} \cdot \left(\frac{m_1}{d}\right)^{\varphi\left(\frac{m_2}{d}\right)-1} \left(\mod \frac{m_2}{d}\right),$$

i.e., if and only if

$$x = c_1 + tm_1 = c_1 + (c_2 - c_1) \cdot \left(\frac{m_1}{d}\right)^{\varphi\left(\frac{m_2}{d}\right)} + r\frac{m_1 m_2}{d}$$
$$= c + r \cdot [m_1, m_2], \text{ where } r \in \mathbb{Z} \text{ is arbitrary and}$$

$c = c_1 + (c_2 - c_1) \cdot (m_1/d)^{\varphi(m_2/d)}$, 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. $\qquad\square$

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 $m_1, , \ldots, m_k \in \mathbb{N}$ be pairwise coprime, $a_1, \ldots, a_k \in \mathbb{Z}$. Then, the system*

$$x \equiv a_1 \pmod{m_1},$$
$$\vdots$$
$$x \equiv a_k \pmod{m_k}$$

has a unique solution modulo $m_1 \cdot m_2 \cdots m_k$.

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:

Solution. We solve the following system of linear congruences in the standard way:

$$c \equiv 3 \pmod 8,$$
$$c \equiv 7 \pmod{17},$$
$$c \equiv -2 \pmod{21},$$

leading to the solution $c \equiv 1027 \pmod{2856}$, which, together with the additional information, implies that there were exactly 3883 gymnasts. $\qquad\square$

11.C.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) $x \equiv 3 \pmod{29}$,
   $x \equiv 5 \pmod{47}$. $\qquad\bigcirc$

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.C.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

$$23941x \equiv 915 \pmod{2^2},$$
$$23941x \equiv 915 \pmod{3^4},$$
$$23941x \equiv 915 \pmod{11}.$$

Solving these congruences separately, we get the following, equivalent system:

$$x \equiv 3 \pmod 4,$$
$$x \equiv -3 \pmod{81},$$
$$x \equiv -4 \pmod{11}.$$

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 $M := m_1 m_2 \cdots m_r$ and $n_i = M/m_i$ for every i, $1 \le i \le r$. Then, for any i, m_i is coprime to n_i, so there is an integer $b_i \in \{1, \ldots, m_i - 1\}$ 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, $1 \le j \le r$, $i \ne j$. Therefore, the wanted solution of the system is the integer

$$x = a_1 b_1 n_1 + a_2 b_2 n_2 + \cdots + a_r b_r n_r.$$

$\qquad\square$

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, $0 \le a < m$, 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., $m = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$, where $p_1, \ldots, p_k$ 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 $f(0)$, $f(1), \ldots, f(104)$. Therefore, we better factor $105 = 3 \cdot 5 \cdot 7$

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. □

11.C.8. Solve the congruence $3446x \equiv 8642 \pmod{208}$. ○

11.C.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. ○

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 is less than their product is then uniquely representable as a k-tuple of remainders (whose values do not exceed the moduli).

11.C.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. □

11.C.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.

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:

$$x \equiv -1 \pmod 3, \qquad x \equiv -1 \pmod 3,$$
$$x \equiv 1 \pmod 5, \quad \text{and} \quad x \equiv 2 \pmod 5,$$
$$x \equiv -2 \pmod 7 \qquad x \equiv -2 \pmod 7.$$

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 $x \equiv 47 \pmod{105}$.

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*

$$x \equiv a \pmod p,$$
$$f(x) \equiv 0 \pmod{p^n}$$

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 $f\left(c_{n-1} + k \cdot p^{n-1}\right) \equiv 0 \pmod{p^n}$. We know that $p^{n-1} \mid f\left(c_{n-1} + k \cdot p^{n-1}\right)$. Now, we use the binomial theorem for $f(x) = a_m x^m + \cdots + a_1 x + a_0$, where $a_0, \ldots, a_m \in \mathbb{Z}$. We have

$$\left(c_{n-1} + k \cdot p^{n-1}\right)^i \equiv c_{n-1}^i + i \cdot c_{n-1}^{i-1} \cdot k p^{n-1} \pmod{p^n},$$

hence

$$f\left(c_{n-1} + k \cdot p^{n-1}\right) \equiv f(c_{n-1}) + k \cdot p^{n-1} f'(c_{n-1}).$$

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

$$(1110110)_2 \equiv (001)_2 + (110)_2 + (110)_2 \equiv 6 \,(\mathrm{mod}\, 2^3 - 1),$$

$$(1110110)_2 \equiv (001)_2 - (110)_2 + (110)_2 \equiv 1 \,(\mathrm{mod}\, 2^3 + 1).$$

A thoughtful reader has surely noticed the similarity with the criteria for divisibility by 9 and 11, which were discussed in paragraph 11.B.9. □

11.C.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 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

$$
\begin{aligned}
(4 + 9s)^4 + 7(4 + 9s) + 4 \equiv{}& 0 \pmod{27}, \\
4^4 + 4 \cdot 4^3 \cdot 9s + 28 + 63s + 4 \equiv{}& 0 \pmod{27}, \\
256 \cdot 9s + 63s \equiv{}& -288 \pmod{27}, \\
256s + 7s \equiv{}& -32 \pmod 3, \\
2s \equiv{}& 1 \pmod 3, \\
s \equiv{}& 2 \pmod 3.
\end{aligned}
$$

Therefore,

$$f\left(c_{n-1} + k \cdot p^{n-1}\right) \equiv 0 \pmod{p^n} \iff$$

$$\iff 0 \equiv \frac{f(c_{n-1})}{p^{n-1}} + k \cdot f'(c_{n-1}) \pmod p.$$

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. □

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.

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 $p - 1$. □

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.C.13. Knowing a primitive root modulo 41 from exercise 11.B.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.C.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.B.28), and the substitution $x \equiv 2^y$ then transforms the congruence to

$$2^{5y} \equiv 2^5 \pmod{11},$$

which is equivalent to the linear congruence

$$5y \equiv 5 \pmod{10},$$
$$y \equiv 1 \pmod{2}.$$

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.C.15. Solve the congruence $x^3 - 3x + 5 \equiv 0 \pmod{105}$. $\bigcirc$

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 12.2.4, 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 are multiples of p. In particular, this applies to the absolute term, which equals $(p-1)!+1$. This proves Wilson's theorem. $\square$

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* iff 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.*

11.C.16. Determine the number of solutions of the congruence

$$x^5 \equiv 534 \pmod{23^2}.$$

Solution. The given congruence is equivalent to $x^5 \equiv 5$ (mod 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). $\qquad\square$

11.C.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 $10x^2 - 15 \equiv 0$ (mod 5), which has five solutions. $\qquad\square$

11.C.18. Other types of congruences. Prove that for any natural number n, the integer

$$111 + 2^{2^{2^{2n-1}}}$$

is divisible by 127.

Solution. We are to prove that the congruence

$$2^{2^{2^{2n-1}}} \equiv -111 \pmod{127}$$

is satisfied for every $n \in \mathbb{N}$. This congruence is equivalent to

$$2^{2^{2^{2n-1}}} \equiv 2^{2^2} \pmod{127}.$$

Since $2^7 = 128 \equiv 1$ (mod 127), the order of 2 modulo 127 equals 7, so the congruence to be proved is (by 11.3.5) equivalent to

$$2^{2^{2n-1}} \equiv 2^2 \pmod 7.$$

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 $0 \le y < \varphi(m)$ such that $x \equiv g^y \pmod m$. Similarly, for a given a, there is a unique $b \in \mathbb{Z}; 0 \le b < \varphi(m)$ 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 11.3.5, 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 $a^{\varphi(m)/d} \equiv 1$ (mod 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 $d \mid b$. $\qquad\square$

Corollary. *If the assumptions of the above theorem hold and, moreover, $(n, \varphi(m)) = 1$, the congruence $x^n \equiv a$ (mod 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 $x^2 \equiv 219$ (mod 383).

Similarly, the order of 2 modulo 7 is 3, which leads to the (again equivalent) congruence

$$2^{2n-1} \equiv 2 \pmod 3,$$
$$(-1)^{2n-1} \equiv -1 \pmod 3,$$

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. □

11.C.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 $n \equiv 5, 6, 10 \pmod{21}$. □

11.C.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.

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, $(-1/5) = (4/5) = 1$, 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 nonresidue 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. □

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 11.3.5) 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

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 $n \equiv 2 \pmod 6$. $\qquad\square$

11.C.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.C.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 $(27, 18) \nmid 3$. $\qquad\square$

Corollary. *(1) Any reduced residue system modulo p contains the same number of quadratic residues as nonresidues.*

(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 $\{-\frac{p-1}{2}, \ldots, -1, 1, \ldots, \frac{p-1}{2}\}$), the quadratic residues are exactly those integers which are congruent to one of $(\pm 1)^2, \ldots, (\pm \frac{p-1}{2})^2$. 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.

(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 ± 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 $p_1, p_2, \ldots, p_\ell$ 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 $p_1, p_2, \ldots, p_\ell$. 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 $2, p_1, p_2, \ldots, p_\ell$ 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*.

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.C.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 Eucliean 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

$$y^2 \equiv -4 \pmod{37}.$$

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. □

11.C.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

$$(x - 12)^2 \equiv 4 \pmod{29}.$$

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) $\left(\frac{q}{p}\right) = \left(\frac{p}{q}\right) \cdot (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$.

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.

$$\left(\frac{79}{101}\right) = \left(\frac{101}{79}\right) \qquad \textit{since } 101 \textit{ is congruent to } 1 \textit{ modulo } 4$$

$$= \left(\frac{22}{79}\right)$$

$$= \left(\frac{2}{79}\right) \cdot \left(\frac{11}{79}\right)$$

$$= \left(\frac{11}{79}\right) \qquad \textit{since } 79 \textit{ is congruent to } -1 \textit{ modulo } 8$$

$$= (-1)\left(\frac{79}{11}\right) \qquad \textit{since } 11 \equiv 79 \equiv 3 \pmod 4$$

$$= (-1)\left(\frac{2}{11}\right) = 1 \qquad \textit{since } 11 \equiv 3 \pmod 8.$$

Many proofs of the the quadratic reciprocity law can be

found in literature[6]. 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.,

$$S = \left\{-\tfrac{p-1}{2}, -\tfrac{p-3}{2}, \ldots, -1, 1, \ldots, \tfrac{p-3}{2}, \tfrac{p-1}{2}\right\}.$$

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

$$1 \cdot a, 2 \cdot a, \ldots, \frac{p-1}{2} \cdot a,$$

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 μ instead of $\mu_p(a)$.

Example. We determine $\mu_p(a)$ for the prime $p = 11$ and the integer $a = 3$.

---

[6]In 2000, F. Lemmermeyer stated 233 proofs – see F. Lemmermeyer, *Reciprocity laws. From Euler to Eisenstein*, Springer. 2000

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.C.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.C.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); 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 $(\pm 1)^2, (\pm 2)^2, (\pm 3)^2, \ldots, (\pm 21)^2$ 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.

Now, the reduced residue system we are interested in is $S = \{-5, \ldots, -1, 1, , \ldots, 5\}$, 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 μ function as $(-1)^{\mu_{11}(3)} = (-1)^2 = 1$.

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 $i \in \{1, 2, \ldots, \frac{p-1}{2}\}$, we set a value $m_i \in \{1, 2, \ldots, \frac{p-1}{2}\}$ so that $i \cdot a \equiv \pm m_i \pmod{p}$. We can easily see that if $k, l \in \{1, 2, \ldots, \frac{p-1}{2}\}$ 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 $\{1, 2, \ldots, \frac{p-1}{2}\}$ and $\{m_1, m_2, \ldots, m_{\frac{p-1}{2}}\}$ coincide, which is also illustrated by the above example. Multiplying the congruences

$$
\begin{aligned}
1 \cdot a &\equiv \pm m_1 \pmod{p}, \\
2 \cdot a &\equiv \pm m_2 \pmod{p}, \\
&\vdots \\
\frac{p-1}{2} \cdot a &\equiv \pm m_{\frac{p-1}{2}} \pmod{p}
\end{aligned}
$$

leads to

$$\tfrac{p-1}{2}! \cdot a^{\frac{p-1}{2}} \equiv (-1)^{\mu} \cdot \tfrac{p-1}{2}! \pmod{p},$$

since there are exactly μ 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

11.C.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 ± 1 or ± 3 modulo 8, so we have (by 11.B.7) that either $p^2 \equiv 1 \pmod{16}$ or $p^2 \equiv 9 \pmod{16}$.

  iii) This is clear from the law of quadratic reciprocity. $\square$

11.C.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 μ tells the number of least (in absolute value) negative remainders of integers in the set
$$\left\{-1, -2, \dots, -\frac{p-1}{2}\right\}.$$
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
$$\left\{1 \cdot 2, 2 \cdot 2, 3 \cdot 2 \dots, \frac{p-1}{2} \cdot 2\right\}.$$
For any $k \in \left\{1, 2, \dots, \frac{p-1}{2}\right\}$, 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
$$\left(\frac{-1}{p}\right) = (-1)^{\mu} = (-1)^{\frac{p-1}{4}} = (-1)^{\frac{p-1}{4} \cdot \frac{p+1}{2}} = (-1)^{\frac{p^2-1}{8}},$$

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 $\left\langle\frac{ak}{p}\right\rangle > \frac{1}{2}$, which is iff 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
$$\left(\frac{a}{p}\right) = (-1)^{\mu_p(a)} = (-1)^{\sum_{k=1}^{\frac{p-1}{2}}\left[\frac{2ak}{p}\right]}.$$

Furthermore, if a is odd, then $a + p$ is even and we get
$$\left(\frac{2a}{p}\right) = \left(\frac{2a+2p}{p}\right) = \left(\frac{4\frac{a+p}{2}}{p}\right) = \left(\frac{2}{p}\right)^2 \cdot \left(\frac{\frac{a+p}{2}}{p}\right)$$
$$= (-1)^{\sum_{k=1}^{\frac{p-1}{2}}\left[\frac{(a+p)k}{p}\right]} = (-1)^{\sum_{k=1}^{\frac{p-1}{2}}\left[\frac{ak}{p}\right]} \cdot (-1)^{\sum_{k=1}^{\frac{p-1}{2}}k}.$$

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
$$(1) \qquad \left(\frac{a}{p}\right) = (-1)^{\sum_{k=1}^{\frac{p-1}{2}}\left[\frac{ak}{p}\right]}.$$

Now, let us consider, for given primes $p \neq q$, the set
$$T = \{q \cdot x;\ x \in \mathbb{Z},\ 1 \le x \le (p-1)/2\} \times$$
$$\times \{p \cdot y;\ y \in \mathbb{Z},\ 1 \le y \le (q-1)/2\}.$$
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.
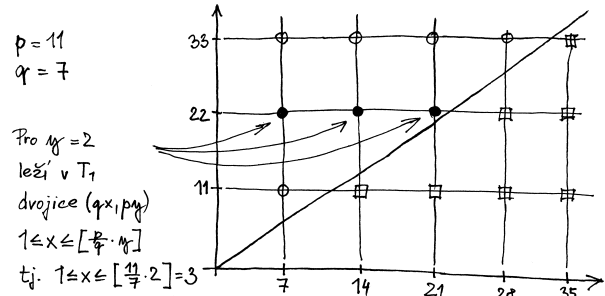
Since the equality $qx = py$ can happen for no pair of x, y from the permissible domain, the set T can be partitioned to (disjoint) subsets T_1 and T_2 so that $T_1 = T \cap \{(u, v);\ u, v \in \mathbb{Z}, u < v\}$, $T_2 = T \setminus T_1$. Clearly, $|T_1|$ is the number of pairs (qx, py) for which $x < \frac{p}{q}y$. Since

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.C.29. Solve the congruence $x^2 - 23 \equiv 0 \pmod{77}$.

Solution. Factoring the modulus, we get the system

$$x^2 - 1 \equiv 0 \pmod{11},$$
$$x^2 - 2 \equiv 0 \pmod 7.$$

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.C.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.C.31. Decide whether the congruence $x^2 \equiv 219 \pmod{383}$ is solvable.

$\frac{p}{q}y \leq \frac{p}{q} \cdot \frac{q-1}{2} < \frac{p}{2}$, 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. □

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) $\left(\frac{a}{b}\right) = \left(\frac{b}{a}\right) \cdot (-1)^{\frac{a-1}{2} \frac{b-1}{2}}$.

PROOF. The proof is simple, utilizing the law of quadratic reciprocity for the Legendre symbol. See exercise 11.C.35. □

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

Solution. Since 383 is a prime, the congruence will be solvable if the Legendre symbol will satisfy $(219/383) = 1$.

$$\left(\frac{219}{383}\right) = -\left(\frac{383}{219}\right) \textit{(Jacobi) as } \textit{383} \equiv \textit{219} \equiv \textit{3} \pmod{\textit{4}}$$

$$= -\left(\frac{164}{219}\right) = -\left(\frac{41}{219}\right) \qquad \textit{164} = \textit{2}^2 \cdot \textit{41}$$

$$= -\left(\frac{219}{41}\right) \quad \textit{(Jacobi) as } \textit{41} \equiv \textit{1} \pmod{\textit{4}}$$

$$= -\left(\frac{14}{41}\right) = -\left(\frac{2}{41}\right)\left(\frac{7}{41}\right)$$

$$= -\left(\frac{7}{41}\right) \qquad \textit{as } \textit{41} \equiv \textit{1} \pmod{\textit{8}}$$

$$= -\left(\frac{41}{7}\right) \qquad \textit{as } \textit{41} \equiv \textit{1} \pmod{\textit{4}}$$

$$= -\left(\frac{-1}{7}\right) = 1 \qquad \textit{as } \textit{7} \equiv \textit{3} \pmod{\textit{4}}.$$

$\square$

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.C.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) $\left(\frac{a}{bb'}\right) = \left(\frac{a}{b}\right)\left(\frac{a}{b'}\right)$.

$\bigcirc$

11.C.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). $\square$

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.*

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 $p \equiv -b^2 \pmod{4q}$. $\square$

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)

11.C.34. Prove that if $a_1, \ldots, a_k$ are odd natural numbers, then

i) $\frac{\prod_{\ell=1}^{k} a_\ell - 1}{2} \equiv \sum_{\ell=1}^{k} \frac{a_\ell - 1}{2} \pmod 2$,

ii) $\frac{\prod_{\ell=1}^{k} a_\ell^2 - 1}{8} \equiv \sum_{\ell=1}^{k} \frac{a_\ell^2 - 1}{8} \pmod 2$.

11.C.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

$$\left(\frac{-1}{a}\right) = \left(\frac{-1}{p_1}\right) \cdot \left(\frac{-1}{p_2}\right) \cdots \left(\frac{-1}{p_k}\right) =$$
$$= (-1)^{\frac{p_1-1}{2}} \cdots (-1)^{\frac{p_k-1}{2}} =$$
$$= (-1)^{\sum_{i=1}^{k} \frac{p_i-1}{2}} =$$
$$= (-1)^{\frac{\prod_{i=1}^{k} p_i - 1}{2}} = (-1)^{\frac{a-1}{2}}.$$

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

$$\left(\frac{p_i}{q_j}\right) = \left(\frac{q_i}{p_j}\right) \cdot (-1)^{\frac{p_i-1}{2} \frac{q_j-1}{2}}.$$

Therefore,

$$\left(\frac{a}{b}\right) = \prod_{i=1}^{k} \prod_{j=1}^{\ell} \left(\frac{p_i}{q_j}\right) =$$
$$= \prod_{i=1}^{k} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) \cdot (-1)^{\frac{p_i-1}{2} \frac{q_j-1}{2}} =$$
$$= \prod_{i=1}^{k} (-1)^{\frac{p_i-1}{2} \sum_{j=1}^{\ell} \frac{q_j-1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) =$$
$$= \prod_{i=1}^{k} (-1)^{\frac{p_i-1}{2} \frac{\prod_{j=1}^{\ell} q_j - 1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) =$$
$$= \prod_{i=1}^{k} (-1)^{\frac{p_i-1}{2} \frac{b-1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) =$$

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.[7]

However, there are cases in which we are able to find the solution of a Diophantine equation, or – at least – to reduce 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

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b,$$

where $x_1, \ldots, x_n$ are unknowns and $a_1, \ldots, a_n, b$ are given non-zero integers.
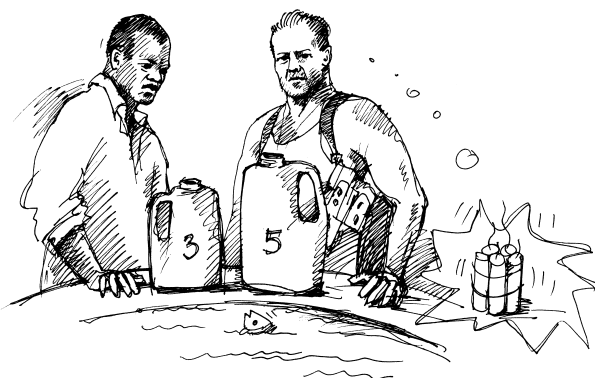
We can see that the ability to solve Diophantine equations is sometimes important in "practical" life as well, as is proved by Bruce Willis and Samuel Jackson in Die Hard with a Vengeance, where they have to do away with a bomb using 4 gallons of water, having only 3- and 5-gallon containers at their 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 $d = (a_1, \ldots, a_n)$ divides b. Provided that, dividing both sides of the equation by the number d leads to an equivalent equation

$$a_1' x_1 + a_2' x_2 + \cdots + a_n' x_n = b',$$

where $a_i' = a_i/d$ for $i = 1, \ldots, n$ and $b' = b/d$. Here, we have

$$d \cdot (a_1', \ldots, a_n') = (da_1', \ldots, da_n') = (a_1, \ldots, a_n) = d,$$

---

[7]See the elementary text M. Davis, *Hilbert's Tenth Problem is Unsolvable*, The American Mathematical Monthly 80(3): 233–269. 1973.

$$= (-1)^{\frac{b-1}{2} \sum_{i=1}^{k} \frac{p_i-1}{2}} \prod_{i=1}^{k} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) =$$

$$= (-1)^{\frac{a-1}{2} \frac{b-1}{2}} \left(\frac{b}{a}\right).$$

We utilized the result of part (i) of the previous exercise in the calculations. □

11.C.36. Determine whether the congruence $x^2 \equiv 38$ (mod 165) is solvable.

Solution. The Jacobi symbol is equal to

$$\left(\tfrac{38}{165}\right) = \left(\tfrac{2}{165}\right) \cdot \left(\tfrac{19}{165}\right) = \left(\tfrac{2}{3}\right) \cdot \left(\tfrac{2}{5}\right) \cdot \left(\tfrac{2}{11}\right)$$
$$\cdot \left(\tfrac{19}{3}\right) \cdot \left(\tfrac{19}{5}\right) \cdot \left(\tfrac{19}{11}\right) = (-1)^3 \left(\tfrac{1}{3}\right) \cdot \left(\tfrac{-1}{5}\right) \cdot \left(\tfrac{2}{11}\right)^3 = 1.$$

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

$$x^2 \equiv -1 \pmod 3,$$
$$x^2 \equiv 3 \pmod 5,$$
$$x^2 \equiv 5 \pmod{11},$$

whence we can easily see that the first and second congruences have no solution. In particular,

$$\left(\tfrac{-1}{3}\right) = -1 \quad \text{and} \quad \left(\tfrac{3}{5}\right) = \left(\tfrac{5}{3}\right) = \left(\tfrac{2}{3}\right) = -1.$$

Therefore, neither the original congruence has a solution. □

11.C.37. Find all primes p such that the integer below is a quadratic residue modulo p:

$$\text{i)}\,3, \quad \text{ii)}\,{-3}, \quad \text{iii)}\,6.$$

Solution.

i) We are looking for primes $p \neq 3$ such that $x^2 \equiv 3$ (mod p) is solvable. Since $p = 2$ satisfies the above, we will consider only odd primes $p \neq 3$ from now on. For $p \equiv 1$ (mod 4), it follows from the law of quadratic reciprocity that $1 = (3/p) = (p/3)$, which occurs if and only if $p \equiv 1$ (mod 3). On the other hand, if $p \equiv -1$ (mod 4), then $1 = (3/p) = -(p/3)$, which holds for $p \equiv -1$ (mod 3). Putting the conditions of both cases together, we arrive at $p \equiv \pm 1$ (mod 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 $p \equiv 1$ (mod 4) and $p \equiv \pm 1$ (mod 12). In the latter case, we must have $p \equiv -1$ (mod 4) and $p \equiv \pm 5$ (mod 12), at the same time – we can take,

so $(a'_1, \ldots, a'_n) = 1$.

Further, we will show that the equation

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b,$$

where $a_1, a_2, \ldots, a_n, b$ are integers such that $(a_1, \ldots, a_n) = 1$, 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 on any parameters). Further, let $n \geq 2$ and suppose that the statement holds for equations having $n-1$ unknowns. Denoting $d = (a_1, \ldots, a_{n-1})$, any n-tuple $x_1, \ldots, x_n$ that satisfies the equation must also satisfy the congruence

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \equiv b \pmod d.$$

Since d is the greatest common divisor of the integers $a_1, \ldots, a_{n-1}$, this congruence is of the form

$$a_n x_n \equiv b \pmod d,$$

which (since $(d, a_n) = (a_1, \ldots, a_n) = 1$) has a unique solution

$$x_n \equiv c \pmod d,$$

where c is a suitable integer, i.e., $x_n = c + d \cdot t$, where $t \in \mathbb{Z}$ is arbitrary. Substituting into the original equation and refining it leads to the equation

$$a_1 x_1 + \cdots + a_{n-1} x_{n-1} = b - a_n c - a_n d t$$

with $n-1$ unknowns and one parameter, t. However, the number $(b - a_n c)/d$ is an integer, so we can divide the equation by d. This leads to

$$a'_1 x_1 + \cdots + a'_{n-1} x_{n-1} = b',$$

where $a'_i = a_i/d$ for $i = 1, \ldots, n - 1$ and $b' = ((b - a_n c)/d) - a_n t$, satisfying

$$(a'_1, \ldots, a'_{n-1}) = (da'_1, \ldots, da'_{n-1}) \cdot \tfrac{1}{d} = (a_1, \ldots, a_{n-1}) \cdot \tfrac{1}{d} = 1.$$

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,

for instance, the set $\{-5, -1, 1, 5\}$ for a reduced residue system modulo 12, and since $(3/p) = 1$ for $p \equiv \pm 1$ (mod 12), we surely have $(3/p) = -1$ whenever $p \equiv \pm 5$ (mod 12). We have thus obtained four systems of two congruences each. Two of them have no solution, and the remaining two are satisfied by $p \equiv 1$ (mod 12) and $p \equiv -5$ (mod 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 $p \equiv \pm 1$ (mod 8) as well as $p \equiv \pm 1$ (mod 12). Solving the corresponding systems of linear congruences leads to the condition $p \equiv \pm 1$ (mod 24). In the latter case, we get $p \equiv \pm 3$ (mod 8) as well as $p \equiv \pm 5$ (mod 12), which together gives $p \equiv \pm 5$ (mod 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. □

11.C.38. The following exercise illustrates that if the modulus of a quadratic congruence is a prime p satisfying $p \equiv 3$ (mod 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 $p \equiv 3$ (mod 4) and an integer a such that $(a/p) = 1$. Prove that the solution of the congruence $x^2 \equiv a$ (mod p) is

$$x \equiv \pm a^{\frac{p+1}{4}} \pmod{p}.$$

Solution. It can be easily verified (using lemma 11.4.12) that
$$\left(a^{\frac{p+1}{4}}\right)^2 \equiv a^{\frac{p+1}{2}} \equiv a \cdot \left(\frac{a}{p}\right) \equiv a \pmod{p}. \qquad \square$$

11.C.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 ($59 \equiv 3$ (mod 4)) that the congruence is satisfied by

$$x \equiv \pm 3^{\frac{59+1}{4}} = \pm 3^{15} \equiv (3^5)^3 \equiv$$
$$\equiv \pm 7^3 = \pm 343 \equiv \mp 11 \pmod{59},$$

since $3^5 = 243 \equiv 7$ (mod 59). □

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 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. □

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

$$(x^2)^2 + (y^2)^2 = z^2,$$

D. 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.D.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:

$$33z \equiv 5 \pmod 7,$$

$$5z \equiv 5 \pmod 7,$$

$$z \equiv 1 \pmod 7.$$

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

$$77x + 63y = 5 - 33(1 + 7t),$$

$$11x + 9y = -4 - 33t.$$

Now, we consider this (parametrized) equation modulo 11:

$$9y \equiv -4 - 33t \pmod{11},$$

$$-2y \equiv -4 \pmod{11},$$

$$y \equiv 2 \pmod{11}.$$

Therefore, this congruence is satisfied by integers $y = 2 + 11s$ for any $s \in \mathbb{Z}$. Now, it only remains to calculate x:

$$11x = -4 - 33t - 9(2 + 11s),$$

$$11x = -22 - 33t - 9 \cdot 11s,$$

$$x = -2 - 3t - 9s.$$

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 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. □

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 $\Theta\left(n^{\log_2 3}\right)$ 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

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

$$mx_n = f(x_1, \ldots, x_{n-1}),$$

where m is a natural number and $f(x_1, \ldots, x_{n-1}) \in \mathbb{Z}[x_1, \ldots, x_{n-1}]$ is a polynomial with integer coefficients, an n-tuple of integers $x_1, \ldots, x_n$ is a solution of it if and only if

$$f(x_1, \ldots, x_{n-1}) \equiv 0 \pmod{m}.$$

11.D.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 fact that the congruence is not solvable can also be established by trying out all four possible remainders modulo 4 into it). $\square$

11.D.3. Solve the following equation in integers:

$$379x + 314y + 183y^2 = 210$$

used, for example, when looking for large primes in the GIMPS project).

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[8] 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 $2^{64} \bmod 1000$, 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.

---

[8]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 (as of July 29, 2017).

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:

$$379 = 2 \cdot 183 + 13,$$
$$183 = 14 \cdot 13 + 1,$$

whence

$$1 = 183 - 14 \cdot 13 = 183 - 14 \cdot (379 - 2 \cdot 183) =$$
$$= 29 \cdot 183 - 14 \cdot 379.$$

Therefore, we can take, for instance, the integer 29 to be our t. Now, multiplying bith 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 $z = y + 5$)

$$(y+5)^2 - 5^2 - 26 \equiv 0 \pmod{379},$$
$$z^2 \equiv 51 \pmod{379}.$$

Invoking the law of quadratic reciprocity, we calculate the Legendre symbol $(51/379)$:

$$\left(\frac{51}{379}\right) = \left(\frac{3}{379}\right) \cdot \left(\frac{17}{379}\right) = \left(\frac{379}{3}\right) \cdot (-1) \cdot \left(\frac{379}{17}\right) \cdot (+1) =$$
$$= \left(\frac{1}{3}\right) \cdot (-1) \cdot \left(\frac{5}{17}\right) = (1) \cdot (-1) \cdot \left(\frac{17}{5}\right) \cdot (+1) =$$
$$= (-1) \cdot \left(\frac{2}{5}\right) = (-1) \cdot (-1) = 1,$$

whence it follows that the congruence is solvable, and, in particular, it has two solutions modulo 379.

The proposition of exercise 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

$$y \equiv 47 \pmod{379}, \qquad y \equiv -57 \pmod{379}.$$

Therefore, the given Diophantine equation is satisfied by those pairs (x, y) with $y \in \{47 + 379 \cdot k;\ k \in \mathbb{Z}\} \cup \{-57 +$

Example. Let us compute $2^{560} \pmod{561}$.

Since $560 = (1000110000)_2$, the mentioned algorithm gives

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.

$379 \cdot k;\ k \in \mathbb{Z}\}$ and $x = \frac{1}{379} \cdot (210 - 314y - 183y^2)$; e. g. $(-1105, 47)$ or $(-1521, -57)$ (which are the only solutions with $|x| < 10^5$). $\qquad\square$

11.D.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 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 $x = 2$, $y = 1$. $\qquad\square$

E. Primality tests
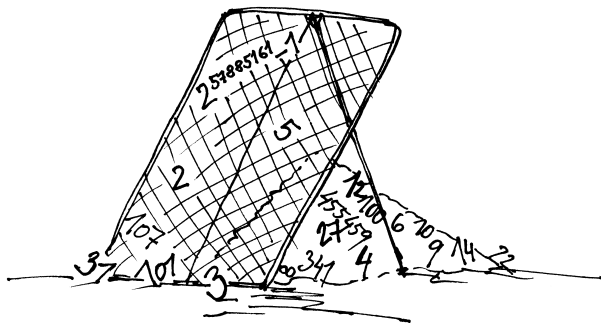
11.E.1. Mersenne primes. The following problems are in deep connection with testing Mersenne numbers for primality.

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

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[9] is $561 = 3 \cdot 11 \cdot 17$, and it was no sooner than in 1992 that it was proved[10] 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. " $\Longleftarrow$ " 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 $i \in \{1, \ldots, k\}$. 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.

" $\Longrightarrow$ " 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 11.3.8, 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 $p_i - 1 \mid n - 1$. $\qquad\square$

---

[9]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.

[10]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.

$q \equiv 3 \pmod 4$ is a Sophie Germain prime[3], 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.

i) If $n \mid q$, then it follows from exercise 11.A.6 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.E.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 $2^{83} - 1$.
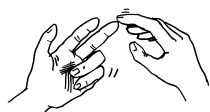
Solution. In the case of the integer $2^{15} - 1$, the exponent is composite; therefore, the whole integer is composite as well (we even know that it is divisible by $2^3 - 1$ and $2^5 - 1$). In the other cases, the exponent is always a prime. We can notice that these primes, namely $q = 11, 23, 29$, 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 $167 \mid 2^{83} - 1$.

[3]See Wikipedia, *Sophie Germain prime*, http://en.wikipedia.org/wiki/Sophie_Germain_prime (as of July 28, 2013, 14:43 GMT).

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.

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 the for N odd, we have $a^{N-1} = (a^{\frac{N-1}{2}} - 1)(a^{\frac{N+1}{2}} - 1)$. □

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. □

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 ± 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[11], 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 $a^q \equiv 1 \pmod p$*

[11]See H. Cohen, *A Course in Computational Algebraic Number Theory*, Springer, 1993.

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

$$p \equiv \pm 1 \pmod 8$$
$$p \equiv 1 \pmod{29},$$

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. $\qquad\square$

11.E.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 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 $\left(\frac{2}{341}\right) = -1$, which follows from the fact that $341 \equiv -3 \pmod 8$. For the integer 561, we have $2^{280} \equiv 1 \pmod{561}$ and $\left(\frac{2}{561}\right) = 1$, 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 $\left(\frac{3}{121}\right) = 1$, but $2^{60} \equiv 89 \not\equiv 1 \pmod{121}$. $\qquad\square$

11.E.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
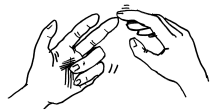
$$a^{n-1} \equiv 1 \pmod n.$$

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 $6600 = 2^3 \cdot 3 \cdot 5^2 \cdot 11$. $\qquad\square$

or there exists an $e \in \{0, 1, \ldots, t-1\}$ such that $a^{2^e q} \equiv -1$ (mod p).

PROOF. It follows from Fermat's little theorem that
$$p \mid a^{p-1} - 1 = (a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1) =$$
$$= (a^{\frac{p-1}{4}} - 1)(a^{\frac{p-1}{4}} + 1)(a^{\frac{p-1}{2}} + 1) =$$
$$\vdots$$
$$= (a^q - 1)(a^q + 1)(a^{2q} + 1) \cdots (a^{2^{t-1}q} + 1),$$

whence the statement follows easily since p is a prime. $\qquad\square$

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*

$$a^q \not\equiv 1 \pmod N$$
$$a^{2^e q} \not\equiv -1 \pmod N \quad for \quad e \in \{0, 1, \ldots, t-1\},$$

then N is not a prime.

PROOF. The correctness of the test follows directly from the previous theorem. $\qquad\square$
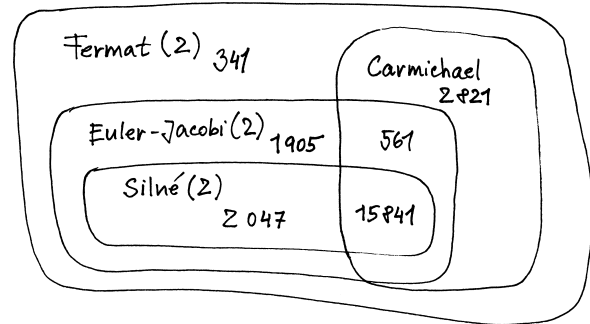
Miscellaneous types of pseudoprimes

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 [12].

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*

11.E.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:

$$2^{952} \equiv -1 \pmod{1905},$$
$$2^{476} \equiv -1 \pmod{1905},$$
$$2^{238} \equiv -1 \pmod{1905},$$
$$2^{119} \equiv \pm 1 \pmod{1905}.$$

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. □

11.E.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 $\left(2^{\frac{1320}{3}} - 1, 1321\right) = 1$ and $\left(2^{\frac{1320}{5}} - 1, 1321\right) = 1$, we can lay $a_3 = a_5 = 2$. However, for $p = 2$, we have $\left(2^{\frac{1320}{2}} - 1, 1321\right) = 1321$, so we have to look for another primality witness. We can take $a_2 = 7$ since $\left(7^{\frac{1320}{2}} - 1, 1321\right) = 1$. In both cases, we have $2^{1320} \equiv$

set $\{a \in \mathbb{Z}; 1 \le a < N, (a, N) = 1\}$ *satisfies the following condition:*

$$a^q \equiv 1 \pmod{N}$$

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[13]).

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 $q \ne 2$ 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 12.2.5). 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) $$s_n = \alpha^{2^n} + \beta^{2^n} = \beta^{2^n}\left(1 + \alpha^{2^{n+1}}\right).$$

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 $s_n = s_{n-1}^2 - 2$

$7^{1320} \equiv 1 \pmod{1321}$. 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. □

11.E.7. Factor the integer 221 to primes by Pollard's ρ-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:

$x := f(x)$	$y := f(f(y))$	$(|x-y|, 221)$ mod 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 $221 = 13 \cdot 17$. □

11.E.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 $455459 = 613 \cdot 743$. □

F. Encryption

11.F.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.

is, by the induction hypothesis, equal to $\left(\alpha^{2^{n-1}} + \beta^{2^{n-1}}\right)^2 - 2 = \alpha^{2^n} + \beta^{2^n}$.

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.B.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).

" $\Longrightarrow$ " 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

$$\alpha^{2^{q-1}} \equiv \left(y\left(1 + \sqrt 3\right)\right)^{M_q+1}$$
$$\equiv y^2 \cdot y^{M_q-1}\left(1 + \sqrt 3\right) \cdot \left(1 + \sqrt 3\right)^{M_q}$$
$$\equiv y^2 \left(1 + \sqrt 3\right) \cdot \left(1 - \sqrt 3\right) = -2y^2$$
$$\equiv -1 \pmod{M_q}.$$

When deriving this, we made use of the fact that 3 is a quadratic nonresidue modulo M_q, so

$$\left(1 + \sqrt 3\right)^{M_q} \equiv 1 + \left(\sqrt 3\right)^{M_q} = 1 + 3^{(M_q-1)/2} \cdot \sqrt 3$$
$$\equiv 1 - \sqrt 3 \pmod{M_q}.$$

" $\Longleftarrow$ " 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 α in the multiplicative group $T_p = \{a + b\sqrt 3; 0 \le a, b < p\} \setminus \{0\}$.

If we had $(3/p) = 1$, then we would get

$$\alpha^{p-1} = \beta \cdot \alpha^p \equiv \beta \cdot \left(2^p + \left(\sqrt 3\right)^p\right)$$
$$\equiv \beta \cdot \left(2 + \sqrt 3 \cdot 3^{(p-1)/2}\right) \equiv \beta \cdot \left(2 + \sqrt 3\right) = 1,$$

whence $p - 1$ would be a multiple of the order of α, i.e. 2^q. However, this would mean that $p > p - 1 \ge 2^q > 2^q -$

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[4] $t = \lceil\sqrt{n}\rceil, t = \lceil\sqrt{n}\rceil + 1, t = \lceil\sqrt{n}\rceil + 2, \ldots$, until $t^2 - n$ is a square (this condition can, of course, be checked efficiently).

11.F.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: $p = t + s = 153001, q = t - s = 151007$. $\square$

---

[4]The symbol $\lceil x \rceil$ denotes the *ceiling* if a real number x, i.e., the it is the integer which satisfies $\lceil x \rceil - 1 < x \leq \lceil x \rceil$.

$1 = M_q$, which contradicts the fact that p is a divisor of M_q. Therefore, we have $(3/p) = -1$ and

$$\alpha^{p+1} \equiv \left(2 + \sqrt{3}\right)\left(2 + \sqrt{3}\right)^p$$
$$\equiv \left(2 + \sqrt{3}\right)\left(2 - \sqrt{3}\right)$$
$$\equiv 1 \pmod{p}.$$

The order of α 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 := 2^q - 1
    repeat  q - 2  times
        s := s^2 - 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.

11.F.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

$$x^2 - (n + 1 - \varphi(n))x + n.$$

11.F.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

$$p = \frac{1}{2}(304008 + \sqrt{304008^2 - 4 \cdot 23104222007} = 153001,$$

$$q = \frac{1}{2}(304008 - \sqrt{304008^2 - 4 \cdot 23104222007} = 151007.$$

$\square$

11.F.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?
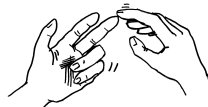
Solution. For completeness, we will first compute the whole public key $A = 9$ (however, this integer was only needed by John when he encrypted the message for Martin; it is no longer necessary for decryption). The message M can be obtained as $M \equiv (6/22^{10}) \pmod{41}$. First, we compute

$$22^{10} \equiv 22^2 \cdot (22^2)^2 \cdot ((22^2)^2)^2 \equiv$$
$$\equiv (-8) \cdot (-8)^2 \cdot (-8)^2 \equiv$$
$$\equiv (-8) \cdot 23 \cdot 23 \equiv -9 \pmod{41}$$

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*

$$a_p^{N-1} \equiv 1 \pmod N \quad and \quad \left(a_p^{\frac{N-1}{p}} - 1, N\right) = 1.$$

Let p^{α_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 $\left(a_p^{(N-1)/p} - 1, N\right) = 1$, 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 $p^{\alpha_p} \mid d - 1$. $\square$

and $(-9)^{-1} \equiv 9 \pmod{41}$. Therefore, the decrypted message is the integer

$$M = 9 \cdot 6 \equiv 13 \pmod{41}. \qquad \square$$

11.F.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

$$r \equiv C^{(p+1)/4} \equiv 692^{\frac{23+1}{4}} \equiv 18 \pmod{23},$$

$$s \equiv C^{(q+1)/4} \equiv 692^{\frac{31+1}{4}} \equiv 14 \pmod{31},$$

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 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. $\qquad \square$

11.F.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, $M = -116 \equiv 321 \pmod{437}$. $\qquad \square$

11.6.12. Theorem. *Let $N \in \mathbb{N}$, $N > 1$. Suppose that we can write $N - 1 = F \cdot U$, where $(F, U) = 1$ and $F > \sqrt{N}$, and that we are familiar with the prime factorization of F. Then:*

- *if we can find for every prime $p \mid F$ an integer $a_p \in \mathbb{Z}$ from the above theorem, then N is a prime;*
- *if N is a prime then for every prime $p \mid N - 1$, there is an integer $a_p \in \mathbb{Z}$ with the desired properties.*

PROOF. By theorem 11.6.11, the potential divisor $d > 1$ of the integer N satisfies $d \equiv 1 \pmod{p^{\alpha_p}}$ for all prime factors of F, hence $d \equiv 1 \pmod{F}$, and so $d > \sqrt{N}$. If N has no non-trivial divisor less than or equal to $\sqrt{N}$, then it is necessarily a prime. On the other hand, it suffices to choose for a_p a primitive root modulo the prime N (independently of p). Then, it follows from Fermat's theorem that $a_p^{N-1} \equiv 1 \pmod{N}$, and since a_p is a primitive root, we get $a_p^{(N-1)/p} \not\equiv 1 \pmod{N}$ for any $p \mid N - 1$.

The integers a_p are again called primality witnesses for the integer N. $\qquad \square$

Remark. The previous test also contains Pépin's test in itself (here, for $N = F_n$, we have $p = 2$, which is satisfied by the primality witness $a_p = 3$).

11.6.13. The polynomial test. Viz Radan (ATC2014) - podrobne, a velmi strucne McAndrew (Crypto in sage) Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni

o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse. Pridat popis AKS algoritmu – zvazit zda vcetne dukazu (podivat se do algebraicke kapitoly, jestli je tam vse potrebne). Pripadne by bylo mozne pridat dukaz tvrzeni o Rabin-Millerovi ze Schoofova clanku citovaneho vyse.

11.6.14. Looking for divisors. If one of the composite-ness 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 compos-iteness, 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 ρ-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.E.8) one of these algorithms – Pollard's ρ-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 $f : S \to S$, 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 $\sqrt{\pi \cdot n/8}$.
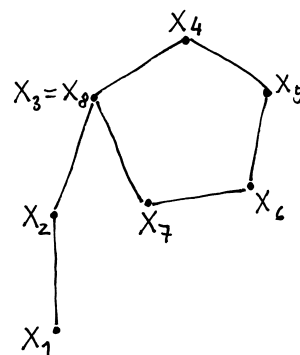
The algorithm described below is again a straightforward implementation of the mentioned reasonings.

Algorithm (Pollard's ρ-method):

```
Input:  n  —  the  integer  to  be  factored ,
         and  an  appropriate  function  f(x)
```

$a := 2; b := 2; d := 1$
While $\quad d = 1$ do
$\quad a := f(a)$
$\quad b := f(f(b))$
$\quad d := \gcd(a - b, n)$
If $\quad d = n$, 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. Its 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 principle of the protocol RSA[14] is as follows:

- Every participant 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 public; the idea is that it is too hard to compute $\varphi(n)$.
- Then, the user chooses a *public key* e and verifies that $(e, \varphi(n)) = 1$.
- Using 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 then 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 for participant A (by any other participant which has access to the public key V_A):

$$C = V_A(M) \equiv M^e \pmod{n}.$$

- Decrypting the cipher C by participant A:

$$OT = S_A(C) \equiv C^d \pmod{n}.$$

---

[14]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 11.3.5, 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.F.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*[15], 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*

---

[15]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.C.38). Indeed, it holds that

$$(\pm x)^2 = (aps + bqr)^2 \equiv (bqr)^2$$
$$\equiv r^2 \equiv C^{(p+1)/2} \equiv C \pmod p,$$

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 use 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 if 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*[16], 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 $(\bmod\ p)$.
- Bob chooses a random integer b and sends g^b $(\bmod\ p)$.
- The common key for the communication is then g^{ab} $(\bmod\ p)$.

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 11.3.7.

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 $h = g^x$ $(\bmod\ p)$, 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 $C_1 = g^y$ $(\bmod\ p)$ and $C_2 = M \cdot h^y$ $(\bmod\ p)$, 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.

---

[16] 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.

G. Additional exercises to the whole chapter

11.G.1. Find all three-digit $n \in \mathbb{N}$, for which $11 \mid n$ a $\frac{n}{11}$ is the sum of squares of the digits of n. ○

11.G.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.G.3. Prove that last four digits of 5^n, $(n = 4, 5, 6, \ldots)$ form a periodical sequence. Find this period. ○

11.G.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.G.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.G.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.G.7. Prove that no natural number n greater than 1 satisfies $n \mid 2^n - 1$. ○

11.G.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.G.9. Let a function $f : \mathbb{N} \to \mathbb{N}$ 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.6 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.G.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

9.B.6. 4π.

9.B.7. 36π.

9.B.8. $\frac{65\pi}{24}$.

10.C.10. $\frac{3}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot 1 = \frac{4}{5}$.

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.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. $\varrho_{U,V} \doteq -0,421$.

11.B.14. Let us consider the factorization of the number n to primes. If $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$, then $\varphi(n) = (p_1 - 1)p_1^{\alpha_1 - 1} \cdots (p_k - 1)p_k^{\alpha_k - 1}$. And if we want to have $9 \mid \varphi(n)$, then at least one of the following conditions must hold:

 i) $p_i \equiv 1 \pmod 9$ for some $i \in \{1, \ldots, k\}$,
 ii) $p_i = 3$ and $\alpha_i \geq 3$ for some $i \in \{1, \ldots, k\}$,
 iii) $p_i = 3$, $\alpha_i = 2$, and $p_j \equiv 1 \pmod 3$ for some distinct $i, j \in \{1, \ldots, k\}$,
 iv) $p_i \equiv 1 \pmod 3$ and $p_j \equiv 1 \pmod 3$ for some distinct $i, j \in \{1, \ldots, k\}$.

If we restrict our attention (as the statement of the problem asks) to numbers $n < 100$, then the condition

 i) is satisfied by primes 19, 37, and 73 (together with their multiples $38, 57, 76, 95$, and 74),
 ii) is satisfied by $3^3 = 27, 3^4 = 81$ (together with a multiple 54),
 iii) is matched by the number $3^2 \cdot 7 = 63$,
 iv) is matched by the number $7 \cdot 13 = 91$.

11.B.22.

 i) The integer 3 has order 4 modulo 10, so it suffices to determine the remainder of the exponent when divided by 4. This remainder is equal to 1, so the last digit is $3^1 = 3$.
 ii) $37 \equiv -3 \pmod{10}$ is of order 4. Again, it suffices to compute the remainder of the exponent upon division by 4. However, we apparently have $37 \equiv 1 \pmod 4$, so the wanted remainder upon division by 10 equals $(-3)^1 \equiv 7$, and the last digit is thus 7.
 iii) Since $(12, 10) > 1$, it makes no sense to talk about the order of 12 modulo 10. However, the examined integer is clearly even, so it suffices to find its remainder upon division by 5. The order of $12 \equiv 2 \pmod 5$ is 4, and the exponent satisfies $13^{14} \equiv 1^{14} \equiv 1 \pmod 4$. We thus have $12^{13^{14}} \equiv 2^1 \pmod 5$, and since 2 is an even integer, it is the wanted digit as well.

11.B.23. Since $\varphi(n) \leq n$, we surely have $\varphi(n) \mid n!$, whence the statement already follows as odd positive integers n satisfy $2^{\varphi(n)} \equiv 1 \pmod n$.

11.B.26. Similarly to the above exercise, we will examine the remainders upon division by coprime integers 125 and 8. We know that $(12, 125) = 1$ and $\varphi(125) = 100$, so

$$12^{10^{11}} \equiv 12^{10^2 \cdot 10^9} = \left(12^{10^2}\right)^{10^9} \equiv 1^{10^9} \equiv 1 \pmod{125}.$$

Since $4 \mid 12$, the number $12^{10^{11}}$ is divisible even by $4^{10^{11}}$, so it is by mere 8 as well, hence $12^{10^{11}} \equiv 0 \pmod 8$. The Chinese remainder theorem states that exactly one of the integers $0, 1, \ldots, 999$ leaves remainder 1 upon division by 125 and is divisible by 8. This integer is 376 (it can be found, for instance, by going through the multiples of 125 increased by 1 and examining their divisibility by 8). Therefore, the last three digits of the number $12^{10^{11}}$ are 376.

11.C.6. i) The greatest common divisor of the moduli is 3, and $1 \not\equiv -1 \pmod 3$, so the system has no solution.

ii) The condition for solvability of linear congruences, $(8, 12345678910111213) = 1$, is clearly true, so this congruence has a unique solution.

iii) The moduli are coprime, so by the Chinese remainder theorem, there is a unique solution modulo $29 \cdot 47$.

11.C.8. We have $208 = 2^4 \cdot 13$ and $(3446, 208) = 2 \mid 8642$. Therefore, the congruence has two solutions modulo 208 and it is equivalent to the system

$$3x \equiv 1 \pmod{8},$$
$$x \equiv 10 \pmod{13}.$$

The solutions of this system are $x \equiv 75$ and $x \equiv 179 \pmod{208}$.

11.C.9. Since 2 is a primitive root modulo both 5 and 13, we get that

$$2^n \equiv 3 \pmod 5 \iff 2^n \equiv 2^3 \pmod 5 \iff n \equiv 3 \pmod 4$$

and

$$2^n \equiv 3 \pmod{13} \iff 2^n \equiv 2^4 \pmod{13} \iff n \equiv 4 \pmod{12}.$$

This apparently implies the infinitude of the multiples of both 5 and 13 among the integers $2^n - 3$ in question. On the other hand, we can see that none of them can be a multiple of 5 and 13 simultaneously since the system of congruences $n \equiv 3 \pmod 4, n \equiv 4 \pmod{12}$ has no solution.

11.C.15. Since the modulus can be written as $105 = 3 \cdot 5 \cdot 7$, where the factors are pairwise coprime, the congruence in question is equivalent to the following system:

$$x^3 - 3x + 5 \equiv 0 \pmod 3,$$
$$x^3 - 3x + 5 \equiv 0 \pmod 5,$$
$$x^3 - 3x + 5 \equiv 0 \pmod 7.$$

Clearly, the first congruence is equivalent to $x^3 \equiv 1 \pmod 3$, and that one is equivalent to $x \equiv 1 \pmod 3$ as it follows from Fermat's little theorem that $x^3 \equiv x \pmod 3$ holds for all integers x.

The second congruence is equivalent to $x(x^2 - 3) \equiv 0 \pmod 5$, which is satisfied iff $x \equiv 0 \pmod 5$ or $x^2 \equiv 3 \pmod 5$. However, since 3 is a quadratic nonresidue modulo 5 (the Legendre symbol $(3/5)$ is equal to -1), we get that $x \equiv 0 \pmod 5$ is the only solution of the second congruence of the system.

The third congruence can be transformed to the form $x^3 - 3x - 2 \equiv 0 \pmod 7$, which is satisfied iff $x \equiv -1 \pmod 7$ or $x \equiv 2 \pmod 7$ (since the left-hand side factors as $x^3 - 3x - 2 = (x-2)(x+1)^2$). Of course, this can also be found out by examining all possibilities modulo 7. Altogether, there are two solutions of the original congruence modulo 105: $x \equiv 55$ and $x \equiv 100$.

11.C.30. The modulus 473 factors to $11 \cdot 43$, thus we have to solve the system of two congruences. The first one leads to $(x-3)^2 \equiv 3 \pmod{11}$ with two solutions $x - 3 \equiv \pm 5 \pmod{11}$. The second one can be transformed to $(x+8)^2 \equiv 15 \pmod{43}$. Now we can proceed either noticing that $15 \equiv 144 \pmod{43}$ or calculating (using result of 11.C.38) $\pm(x+8) \equiv 15^{\frac{43+1}{4}} \equiv 31 \pmod{43}$, in both cases we get the result $x \equiv 4, 23 \pmod{43}$. Now we combine the solutions of both congruences of the system and we obtain $x \equiv 152, 195, 262, 305 \pmod{473}$.

11.C.32. All of the results can be proved directly from the definition of the Jacobi symbol and the multiplicativity of the Legendre symbol.

11.C.34. In light of the previous exercise, both statements can be proved easily by mathematical induction.

11.G.1. Let $n = \overline{abc}$, where $a \ne 0$. Then $11 \mid n \iff 11 \mid c + a - b \iff c + a - b \in \{0, 11\}$.
We should have $\frac{100a + 10b + c}{11} = a^2 + b^2 + c^2 \iff 100a + 10b + c = 11(a^2 + b^2 + c^2)$.
If $c + a - b = 0$, i.e. $b = a + c$, then

$$100a + 10(a+c) + c = 11(a^2 + (a+c)^2 + c^2)$$
$$110a + 11c = 11(2a^2 + 2ac + 2c^2)$$
$$10a + c = 2a^2 + 2ac + 2c^2$$
$$2a^2 + 2ac - 10a + 2c^2 - c = 0$$
$$a^2 + (c-5)a + c^2 - \frac{c}{2} = 0.$$

The discriminant of this quadratic equation is $-3c^2 - 8c + 25 \geq 0 \iff c \in \{0, 1\}$.
$c = 0 \implies a^2 - 5a = 0 \implies a = 5, b = 5$. Thus the first solution is $n = 550$.
$c = 1 \implies a, b \notin \mathbb{N}$.

If $c + a - b = 11$, i.e. $b = a + c - 11$, then

$$100a + 10(a + c - 11) + c = 11(a^2 + (a + c - 11)^2 + c^2)$$
$$110a + 11c - 110 = 11(2a^2 + 2ac + 2c^2 - 22a - 22c + 121)$$
$$10a + c - 10 = 2a^2 + 2ac + 2c^2 - 22a - 22c + 121$$

$$2a^2 + 2ac - 32a + 2c^2 - 23c + 131 = 0$$

$$a^2 + (c - 16)a + c^2 - \frac{23}{2}c + \frac{131}{2} = 0.$$

Now, the discriminant is $-3c^2 + 14c - 6 \geq 0 \iff c \in \{1, 2, 3, 4\}$.

$c = 1 \vee c = 2 \vee c = 4 \implies a, b \notin \mathbb{N}$.

$c = 3 \implies a^2 - 13a + 40 = 0 \implies a = 8, b = 0$. The other solution is therefore $n = 803$.

11.G.2. Fermat's little theorem states that $a^p \equiv a$ which together with the requirement $a^p \equiv 1$ gives the condition $a \equiv 1 \pmod{p}$ for the pairs (a, p).

For $a < 16$ we get the following pairs:

a	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p	p	–	2	3	2	5	2,3	7	2	3	2,5	11	2,3	13	2,7

11.G.3. We prove by the induction that $5^n \equiv 5^{n+4} \pmod{10000}$ for $n \geq 4$.

For $n = 4$ we have $5^4 = 625 \equiv 390625 = 5^8 \pmod{10000}$.

Induction step: $5^{k+1} \equiv 5^{k+5} \equiv 5^{k+4} \cdot 5 \equiv 5^k \cdot 5 \equiv 5^{k+1} \pmod{10000}$.

For all $a \in \mathbb{N}$ we therefore have: $5^{4a} \equiv 625 \pmod{10000}$, $5^{4a+1} \equiv 3125 \pmod{10000}$, $5^{4a+2} \equiv 5625 \pmod{10000}$ a $5^{4a+3} \equiv 8125 \pmod{10000}$, and the last four digits of 5^n form periodical sequence with period 4.

11.G.4. If n is an arbitrary natural number, then $2^{2^n} \equiv 1 \pmod{3}$, so it suffices to choose for k odd positive integers with $k \equiv 2 \pmod{3}$. And there are surely infinitely many of them – they are those which satisfy $k \equiv 5 \pmod{6}$. For these values of k, we always have that $2^{2^n} + k$ is a multiple of 3 and greater than 3, so it is a composite number.

11.G.5. Let us fix an integer $k \in \mathbb{Z} \setminus \{1\}$ and an arbitrary $a \in \mathbb{N}$. We will show that for an arbitrarily large a, we can find a positive integer n such that $2^{2^n} + k$ is composite and greater than a. That will complete the proof.

Let us fix $s \in \mathbb{N}_0, h \in \mathbb{Z}$ so that $k - 1 = 2^s \cdot h, 2 \nmid h$, and $m \in \mathbb{N}$ satisfying $2^{2^m} > a - k$. Now, let an ℓ satisfy $\ell \geq s, \ell \geq m$. If the integer $2^{2^\ell} + k$ is composite, then we are done, since $2^{2^\ell} + k \geq 2^{2^m} + k > a$. Therefore, let us assume that the integer $2^{2^\ell} + k$ is a prime and denote it by p. With help of Euler's theorem, we can find an integer of the desired form which is a multiple of p. We have

$$p - 1 = 2^{2^\ell} + 2^s \cdot h = 2^s \cdot h_1,$$

where $h_1 \in \mathbb{N}$ is odd. We thus have $2^{\varphi(h_1)} \equiv 1 \pmod{h_1}$, whence $2^{s+\varphi(h_1)} \equiv 2^s \pmod{p - 1}$, and since $l \geq s$, we also have

$$2^{\ell + \varphi(h_1)} \equiv 2^\ell \pmod{p - 1}.$$

Now, it follows from Fermat's little theorem that

$$2^{2^{\ell + \varphi(h_1)}} + k \equiv 2^{2^\ell} + k \equiv 0 \pmod{p}.$$

However, since $2^{\ell + \varphi(h_1)} > 2^\ell$, we also have $2^{2^{\ell + \varphi(h_1)}} + k > 2^{2^\ell} + k = p > a$. We have thus found a composite number which is of the wanted form and greater than an arbitrarily large value of a.

Let us mention that the case of $k = 1$ is a well-known open problem examining the infinitude of Fermat primes.

11.G.6. We can easily see that $2 \mid a_1 = 10$ and $3 \mid a_2 = 48$. Further, we can show that $p \mid a_{p-2}$ holds for any prime $p > 3$. By Fermat's theorem, we have $2^{p-1} \equiv 3^{p-1} \equiv 6^{p-1} \equiv 1 \pmod{p}$. Therefore,

$$6a_{p-2} = 3 \cdot 2^{p-1} + 2 \cdot 3^{p-1} + 6^{p-1} - 6 \equiv 3 + 2 + 1 - 6 = 0 \pmod{p}.$$

Let us remark that knowledge of algebra allows us to proceed more directly: for $p > 3$, we can consider the p-element field $\mathbb{F}_p$, which contains multiplicative inverses of the elements 2, 3, and 6 and their sum is $\frac{1}{2} + \frac{1}{3} + \frac{1}{6} = 1$.

11.G.7. We could reason about the factorization of n to primes, which is a bit complicated. Instead, we will use a little trick. Suppose that there is an n satisfying the conditions $n \mid 2^n - 1, n > 1$, and let us select the least one. Surely, n is odd, hence $n \mid 2^{\varphi(n)} - 1$. Utilizing the result of exercise , we get that $n \mid 2^d - 1$, where $d = (n, \varphi(n))$ (which especially implies that $2^d - 1 > 1$ and $d > 1$). At the same time, $d \leq \varphi(n) < n$ and $d \mid n$, whence it follows that $d \mid 2^d - 1$, which contradicts the assumption that our n is the least one that meets the conditions.

11.G.8. Since $2^{p-1} \equiv 1 \pmod{p}$, it suffices to choose appropriate multiples of $p-1$ for n, i. e., to find a k so that $n = k(p-1)$ would satisfy the condition $n \cdot 2^n \equiv -1 \pmod{p}$. However, thanks to $p-1 \mid n$, this is equivalent to $k \equiv 1 \pmod{p}$, and there are clearly infinitely many such values k.

11.G.9. Analyze the Euclidean algorithm for computing the greatest common divisor.

11.G.10. The signature is $8^{103} \equiv 83 \pmod{143}$, which can be verified by $83^7 \equiv 8 \pmod{143}$. Finally, 42 is not the valid signature as $42^7 \equiv 81 \not\equiv 26 \pmod{143}$.

Algebraic structures

The more abstraction, the more chaos?
– no, it is often the other way round...

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.

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 φ.

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 $\vee : K \times K \to K$ and intersection $\wedge :$

A	B	C	$B' \Rightarrow C$	$[(A \vee C) \wedge B]'$	φ
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:

$$\begin{aligned}
&(B' \Rightarrow C) \wedge [(A \vee C) \wedge B]' \Longleftrightarrow \\
\Longleftrightarrow &(B \vee C) \wedge [(A \vee C)' \vee B'] \Longleftrightarrow \\
\Longleftrightarrow &(B \vee C) \wedge [(A' \wedge C') \vee B'] \Longleftrightarrow \\
\Longleftrightarrow &[(B \vee C) \wedge (A' \wedge C')] \vee [(B \vee C) \wedge B'] \Longleftrightarrow \\
\Longleftrightarrow &[(B \wedge A' \wedge C') \vee (C \wedge A' \wedge C')] \vee [(B \wedge B') \vee (C \wedge B')] \Longleftrightarrow \\
\Longleftrightarrow &(B \wedge A' \wedge C') \vee (C \wedge B'),
\end{aligned}$$

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*). □

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: $\wedge, \vee, \Longrightarrow, \equiv$ 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 ($A \operatorname{NAND} B = (A \wedge B)'$,

$K \times K \to K$. 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 $' : K \to K$ with only one argument. Such operations are called *unary operations*.

In general, there are algebraic structures with k operations $\mu_1, \ldots, \mu_k$, each of them

$$\mu_j : K \overset{i_j\text{-times}}{\times \cdots \times} K \to K$$

with i_j arguments, and write $(K, \mu_1, \ldots, \mu_k)$ 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:

<div align="center">

AXIOMS OF BOOLEAN ALGEBRAS

</div>

(1)	$A \wedge (B \wedge C) = (A \wedge B) \wedge C,$
(2)	$A \vee (B \vee C) = (A \vee B) \vee C,$
(3)	$A \wedge B = B \wedge A, \ A \vee B = B \vee A,$
(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.

A 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. ○

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

$$((A \wedge B) \vee C)' = (A \wedge B)' \wedge C' = A' \wedge C' \vee B' \wedge C'.$$

Setting 1 for the value *True* and 0 for *False* for A, B, C we obtain the table for $\varphi(A, B, C) = A' \wedge C' \vee B' \wedge C'$:

A	B	C	φ
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 $A = 0, \quad B = 0, \quad C = 1$ 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 swich a is closed we set $a = 1$ and so on. Switches labeled with the same letter must are 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 Bolean proposition.

Properties (6)–(8) require the existence of neutral elements for both operations as well as the existence of an analogy 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.

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

$$B = B \vee 0 = B \vee (A \wedge C) =$$
$$= (B \vee A) \wedge (B \vee C) = 1 \wedge (B \vee C) = B \vee C$$

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*.

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,

$$\varphi = a \wedge (b \vee d' \vee c \wedge (a \vee d \vee c')) \wedge b.$$

$\square$

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??? $\square$

12.A.8. Verify equivalence of the following two Boolean propositions

$(a \vee b) \wedge (c \vee d)$ and $(a \wedge c) \vee (a \wedge d) \vee (b \wedge c) \vee (b \wedge d)$

Solution. Distributive laws yield

$(a \vee b) \wedge (c \vee d) = ((a \vee b) \wedge c) \vee ((a \vee b) \wedge d) = a \wedge c \vee b \wedge c \vee a \wedge d \vee b \wedge d.$

$\square$

12.A.9. Determine whether the following two Boolean propositions

$$((a' \wedge b) \vee (a \wedge c'))'   \text{and}   (a \vee b') \wedge (a \vee c')$$

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. $\square$

12.A.10. Determine whether the following two Boolean propositions

$$(a \wedge b \wedge c) \vee (a \vee c)'   \text{and}   (a \wedge c) \vee (a' \wedge c')$$

are equivalent?

Solution. setting $a = 1, b = 0, c = 1$ implies $(a \wedge b \wedge c) \vee (a \vee c)' = 0$ and $(a \vee b') \wedge (a \wedge c) \vee (a' \wedge c') = 1$. Hence, these propositions are not equivalent. $\square$

Proposition. *In every Boolean algebra* $(K, \wedge, \vee)$;
(1) $A \wedge 0 = 0,   A \vee 1 = 1,$
(2) $A \wedge (A \vee B) = A,   A \vee (A \wedge B) = A,$
(3) $A \wedge A = A,   A \vee A = A,$
(4) $(A \wedge B)' = A' \vee B',   (A \vee B)' = A' \wedge B',$
(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):

$$A \wedge (A \vee B) = (A \vee 0) \wedge (A \vee B) =$$
$$= A \vee (0 \wedge B) = A \vee 0 = A.$$

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

$$(A \wedge B) \wedge (A' \vee B') = ((A \wedge B) \wedge A') \vee ((A \wedge B) \wedge B')$$
$$= (0 \wedge B) \vee (A \wedge 0) = 0.$$

Similarly,

$$(A \wedge B) \vee (A' \vee B') = (A \vee (A' \vee B')) \wedge (B \vee (A' \vee B'))$$
$$= (1 \vee B') \wedge (1 \vee A') = 1.$$

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 $A = (A')'$. $\square$

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 $S = \{f : M \to K\}$ 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.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. $\qquad \square$

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

$$(a \cdot b + a' + b) \cdot (b + c + b \cdot c') = \cdots = a' \cdot c + b.$$

This means that the given formula is equivalent to $(A' \wedge C) \vee B$. $\qquad \square$

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 dijunsctive normal form is quite suitable. Consider the proposition:

$$a \wedge b \wedge c' \vee a \wedge b' \wedge c \vee a' \wedge b \wedge c \vee a \wedge b \wedge c.$$

$\qquad \square$

12.A.14. Let A be a free boolean algebra, generated by countably many generators $a_1, \ldots, a_n, \ldots$ 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 $a_1, \ldots, a_n$. Then $f \wedge a_{n+1}$ is neither f nor 0, so f can not be an atom in A. $\qquad \square$

as functions of an argument $x \in M$ as follows:

$$(f_1 \wedge f_2)(x) = (f_1(x)) \wedge (f_2(x)) \in K,$$
$$(f_1 \vee f_2)(x) = (f_1(x)) \vee (f_2(x)) \in K,$$
$$(1)(x) = 1 \in K, \ (0)(x) = 0 \in K,$$
$$(f)'(x) = (f(x))' \in K.$$

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 $M \to \mathbb{Z}_2$ (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.

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 $A, B, \ldots \in K$, 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 $A, B, \ldots$ 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 $(\mathbb{Z}_2)^n \to \mathbb{Z}_2$. 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:

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

$$(A \vee B) \vee (\tilde{A} \vee \tilde{B}) = (A \vee \tilde{A}) \vee (B\tilde{B}) = \tilde{A} \vee \tilde{B},$$

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. □

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 $\wedge$,

$$C \wedge A \leq C \wedge (C' + B) \leq C \wedge B \leq B.$$

□

12.A.17. For any two elements A, B of a Boolean algebra, the *symmetric difference* of A and B is defined as

$$A \bigtriangleup B = A \wedge B' \vee B \wedge A'.$$

This is analogous of taking symmetric difference $A \bigtriangleup B = (A \setminus B) \cup (B \setminus A)$. Show that

i) $A = B$ if and only if $A \bigtriangleup B = 0$.

ii) $A \bigtriangleup (B \bigtriangleup C) = (A \bigtriangleup B) \bigtriangleup C$

iii) $A \wedge (B \bigtriangleup C) = (A \wedge B) \bigtriangleup (A \wedge C)$.

Solution.

(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).
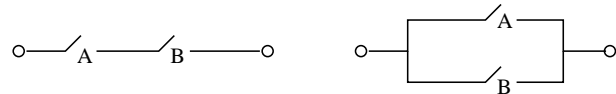
One or more switches may be interconnected in a series circuit or a parallel circuit. The series circuit corresponds to the binary operation $\wedge$, while the parallel circuit corresponds to $\vee$. The unary operation A' defines a switch whose state is always the opposite than that of A. Every finite word created from the switches $A, B, \ldots$ 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.
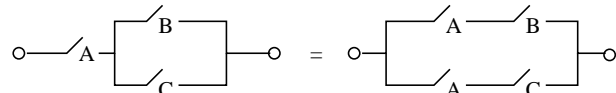
The circuit without a switch corresponds to 1. When the endpoints are not connected, this corresponds to 0 (consider a series circuit of A and A'). Draw diagrams for all axioms of Boolean algebras and verify them!

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

i) If $A = B$ then $A \leq B$ ans $B \leq A$, and therefore, by previous exercise, $A \wedge B' = 0$ and $B \wedge A' = 0$, thus $A \bigtriangleup B = 0$. On the other hand, $A \bigtriangleup B = 0$ implies that both $A \wedge B' = 0$ and $B \wedge A' = 0$, which yield $A \leq B$ ans $B \leq A$, thus $A = B$.

ii) Expanding $(B \bigtriangleup C)'$ using distributivity and de Morgan's laws, we obtain

$$(B \bigtriangleup C)' = (B' \vee C) \wedge (C' \vee B) = B \wedge C \vee B' \wedge C'.$$

Further expanding $A \bigtriangleup (B \bigtriangleup C)$ we get

$$A \bigtriangleup (B \bigtriangleup C) = A \wedge (B \wedge C \vee B' \wedge C') \vee A' \wedge (B \wedge C' \vee B' \wedge C)$$
$$= A \wedge B \wedge C \vee A \wedge B' \wedge C' \vee A' \wedge B \wedge C' \vee A' \wedge B' \wedge C,$$

which equals $(A \bigtriangleup B) \bigtriangleup C$ by obvious symmetry.

iii) Follows from

$$A \wedge (B \bigtriangleup C) = A \wedge (B \wedge C' \vee C \wedge B') = A \wedge B \wedge C' \vee A \wedge \mathbb{C} \wedge B'$$

and

$$(A \wedge B) \bigtriangleup (A \wedge C) = A \wedge B \wedge (A' \vee C') \vee A \wedge C \wedge (A' \vee B') = A \wedge B \wedge C' \vee A \wedge C \wedge B'$$ $\square$

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. $\square$

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. $\square$

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 $p = q_1 \ldots q_n$, 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 $a = q_1^{n_1} \ldots q_s^{n_s}$, $b = q_1^{m_1} \ldots q_s^{m_s}$, $c = q_1^{k_1} \ldots q_s^{k_s}$. 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 a 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). $\square$

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.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 $\to$ mean $A \to B = A' \vee B$. Prove the equivalence of the following logical propositions

  i) $(A' \to (B \to C)) = (B \to (A \vee C))$;

  ii) $(A \to B')' \wedge (A \vee B)' = 0$, that is, it is always *False*.

  iii) $(A \vee B) \wedge (B' \vee C) \to (B \vee C) = 1$ that is, it is always *True*

12.A.22. Show that the following dvisibilty 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 transityive. 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

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. □

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

i) If $\leq$ is a partial order 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, $a < c$. $\square$

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 $a_3, \ldots, a_n$ 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.

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

Solution. Let x' and z denote complements of $x \in A$. Then

(1) $x' = x' \wedge 1$
$$= x' \wedge (x \vee z)$$
$$= (x' \wedge x) \vee (x' \wedge z)$$
$$= 0 \vee (x' \wedge z)$$
$$= (x' \wedge z) \vee (x' \wedge z)$$
$$= (x \vee x') \wedge z$$
$$= 1 \wedge z = z$$

$\square$

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. $\square$

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.

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 842. There are lattices which are not distributive; see the Hasse diagrams of two such simple lattices below (and check that in both cases $x \wedge (y \vee z) \neq (x \wedge y) \vee (x \wedge z)$).
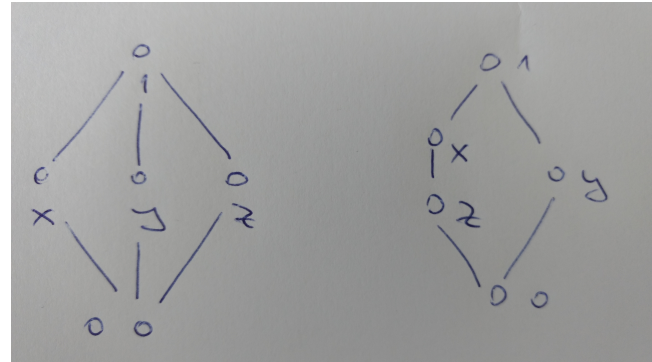
Now, Boolean algebras can be defined in terms of lattices: *a Boolean algebra is a complete distributive lattice such that each element has its complement* (i.e. the axiom 12.1.2(8) is satisfied).

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 $f : K \to L$ 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.

□
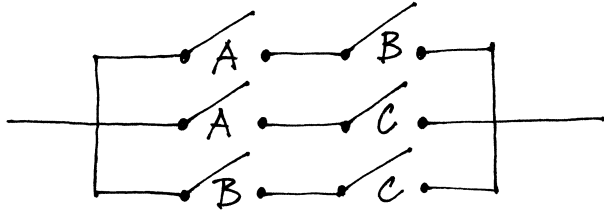
12.A.31. Find a finite subset of the set of positive integers which is not a lattice with respect to divisibility. ○

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:

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.

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. □

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. ○

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 distributive (it does not contain a sublattice isomorphic to diamond or pentagon).

In the case of Boolean algebras, homomorphisms are defined as follows:

LATTICE AND BOOLEAN-ALGEBRA HOMOMORPHISMS

A mapping $f : (K, \wedge, \vee) \to (L, \wedge, \vee)$ 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 $f : K \to L$ 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 $f : K \to K$ 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[1]:

TARSKI'S FIXED-POINT THEOREM

Theorem. *Let* $(K, \wedge, \vee)$ *be a complete lattice and* $f : K \to K$ *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.

---

[1]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.

dělitelé 36

(Hasse diagram with nodes: 1 at bottom; 2, 3; 4, 6, 9; 12, 18; 36 at top)
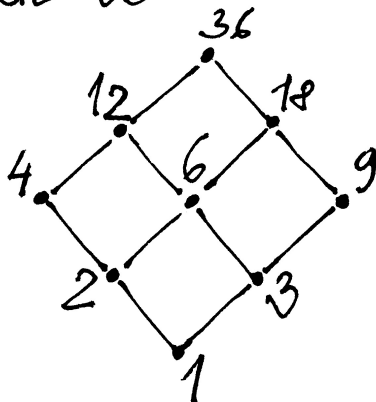
12.A.35. Draw the Hasse diagram of the lattice of all (positive) divisors 30. Is this lattice distributive? Is it a Boolean algebra?

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.

dělitelé 30

(Hasse diagram cube with nodes: 1 at bottom; 2, 3, 5; 6, 10, 15; 30 at top)
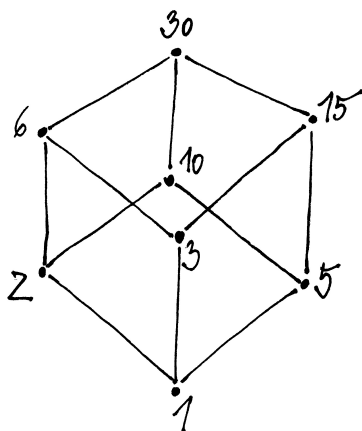
12.A.36. Decide whether every lattice on a 3-element set is a *chain*, i. e., whether each pair of elements are necessary comparable.

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,

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 $f_{|R} : R \to R$. 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*

$$0 \leq f(0) \leq f(f(0)) \leq \ldots.$$

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.[2] 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.

_____

[2]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.

which means that the remaining one is comparable with these two; and there are no other elements. □

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:

□
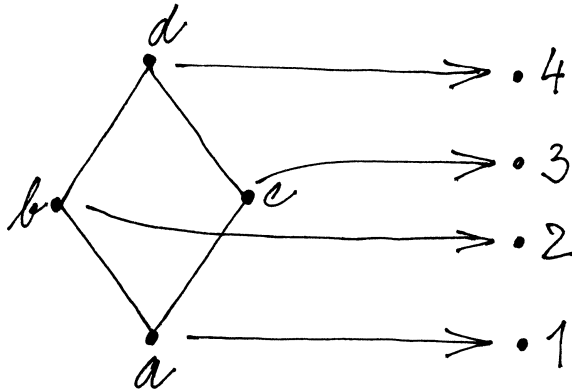
12.A.38. Decide whether every lattice homomorphism between finite non-empty lattices K, L maps the least element of K to the least element of L.

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. □

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:

$$a \prec b \iff [(\text{sgn}(a) \cdot \text{sgn}(b) = 1 \wedge b > a) \vee (\text{sgn}(a) > \text{sgn}(b))].$$

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 $(\mathbb{Z}_2)^n \to \mathbb{Z}_2$ 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 $(\mathbb{Z}_2)^n \to \mathbb{Z}_2$. 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 $(\mathbb{Z}_2)^n \to \mathbb{Z}_2$ 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 $(\mathbb{Z}_2)^n \to \mathbb{Z}_2$ 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.
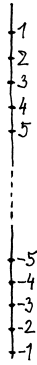
12.A.40. Give an example of an infinite chain which is a complete lattice.

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 ∞ 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 ∞ is the supremum of subsets which are not bounded from above. $\square$

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,

$$K_1 \vee (K_2 \wedge K_3) = K_1 \neq (K_1 \vee K_3) \wedge (K_1 \vee K_2).$$

$\square$

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?

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 $A_1, \ldots, A_n$. It can be easily verified that there are 2^n atoms, which are of the form $A_1^{\sigma_1} \wedge \cdots \wedge A_n^{\sigma_n}$, where either $A_i^{\sigma_i} = A_i$ or $A_i^{\sigma_i} = A_i'$.

The infimum $\varphi \wedge \psi$ of functions φ and ψ is the function whose values are given by the products of the corresponding values in $\mathbb{Z}_2$. Therefore, $\varphi \leq \psi$ if φ takes the value $1 \in \mathbb{Z}_2$ only on arguments where ψ also has value 1. Hence in the Boolean algebra of truth-value functions, a function φ is an atom if and only if φ 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*

$$B = A_1 \vee \cdots \vee A_k.$$

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 $A_1, A_2, \ldots, A_k$ 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

$$Y = A_1 \vee \cdots \vee A_k \leq B.$$

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 $Y, X_1, \ldots, X_\ell$ are atoms in K, then $Y \leq X_1 \vee \ldots \vee X_\ell$ if and only if $Y = X_i$ for some $i = 1, \ldots, \ell$.*
(2) For each $Y \in K$, $Y \neq 0$, there is an atom $X \in K$ such that $X \leq Y$.
(3) If $X_1, \ldots, X_r$ are precisely all the atoms of K, then $Y = 0$ if and only if $Y \wedge X_i = 0$ for all $i = 1, \ldots, r$.

PROOF. (1) If the inequality of the proposition holds, then

$$Y \wedge (X_1 \vee \cdots \vee X_\ell) = Y.$$

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). □

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

By distributivity, the equality can be rewritten as

$$(Y \wedge X_1) \vee \cdots \vee (Y \wedge X_\ell) = Y.$$

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

$$\cdots \leq Z_k \leq Z_{k-1} \leq \cdots \leq Z_1 \leq Y,$$

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

$$Y = A_1 \vee \cdots \vee A_k \leq B,$$

where A_i are all the atoms in K which are less then or equal to B. Compute

$$B \wedge Y' = B \wedge (A_1 \vee \cdots \vee A_k)' = B \wedge A_1' \wedge \cdots \wedge A_k'.$$

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

$$B = A_1 \vee \cdots \vee A_k = \tilde{A}_1 \vee \cdots \vee \tilde{A}_\ell.$$

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

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 φ is a homomorphism with respect to multiplication. Since addition is defined componentwise, φ is a homomorphism to it as well. Moreover, this mapping is clearly both injective and surjective, thus it is an isomorphism. $\square$

12.B.4. Prove that the identity is the only automorphism of the field of real numbers.

Solution. Consider an automorphism $\varphi : \mathbb{R} \to \mathbb{R}$. Clearly, it must satisfy $\varphi(0) = 0$ and $\varphi(1) = 1$. Since φ respects addition, we must have for all positive integers n that $\varphi(n) = \varphi(1 + 1 + \cdots + 1) = n\varphi(1) = n$ and $\varphi(-n) = -n$. Since it respects multiplication, we must have for any integers 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, $\varphi(x) = \varphi\left(\sqrt{x}^2\right) = \varphi\left(\sqrt{x}\right)^2 \geq\geq 0$. Thus, for any $x, y \in \mathbb{R}$ such that $x < y$, we must have $\varphi(x) < \varphi(y)$. Now, assume that φ 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 12.3.8 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).

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:

$$X = A_1 \vee \cdots \vee A_k$$

Define the function $f : K \to 2^M$ by

$$f(X) = f(A_1) \cup \cdots \cup f(A_k) = \{A_1, \ldots, A_k\},$$

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. $\square$

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*[3].

2. 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.

---

[3]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.

Solution. Applying the Euclidean algorithm, we get

$$131 = 7 \cdot 17 + 12,$$
$$17 = 1 \cdot 12 + 5,$$
$$12 = 2 \cdot 5 + 2,$$
$$5 = 2 \cdot 2 + 1.$$

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 $[19]^{-1} = 69$. □

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

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

be a polynomial over $\mathbb{Z}$ and p be a prime such that

- p divides a_j, $j = 0, \ldots, n-1$,
- 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

Among others, the abstract algebraic theory can be in many aspects viewed as a straightforward generalization of divisibility properties of integers.

12.2.1. 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*.

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.

Recall the useful example of a non-commutative ring, the set $\mathrm{Mat}_k(\mathbb{K})$ 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

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.

$$2x^4 + 4x^3 + 6x^2 + 4x + 2 : 2x^3 + 3x^2 + 3x + 1 = x + \frac{1}{2}$$
$$\underline{2x^4 + 3x^3 + 3x^2 + x}$$
$$\quad x^3 + 3x^2 + 3x + 2$$
$$\underline{x^3 + \frac{3}{2}x^2 + \frac{3}{2}x + \frac{1}{2}}$$
$$\qquad \frac{3}{2}x^2 + \frac{3}{2}x + \frac{3}{2}$$

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 $\frac{2}{3}$)

$$2x^3 + 3x^2 + 3x + 1 : x^2 + x + 1 = 2x + 1$$
$$\underline{2x^3 + 2x^2 + 2x}$$
$$\qquad x^2 + x + 1$$

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):

$$x^4 + 2x^3 + 3x^2 + 2x + 1 =$$
$$= \left(x + \frac{1}{2} - i\frac{\sqrt{3}}{2} \right)^2 \cdot \left(x + \frac{1}{2} + i\frac{\sqrt{3}}{2} \right)^2.$$

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!):

$$x^4 + 2x^3 + 3x^2 + 2x + 1 = \left(x^2 + x + 1 \right)^2.$$

- Let us solve the equation

$$x^4 + 2x^3 + 3x^2 + 2x + 1 = 0.$$

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

never an integral domain (see 2.1.5 on page 72 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!

12.2.2. 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 $c = 1 \cdot c = 0 \cdot c = 0$. □

12.2.3. Polynomials over rings. The definition of a commutative 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:[4]

---

[4] 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.

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 $\mathbb{R}$:

$$\left(x^2-\sqrt{2}x+1\right)\left(x^2+\sqrt{2}x+1\right)..$$ □

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 nonzero 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 $\deg R(x)<\operatorname{st}Q(x)$ 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

---

Definition. Let $\mathbb{K}$ be a commutative ring. A polynomial over $\mathbb{K}$ is a finite expression

$$f(x)=\sum_{i=0}^{k}a_ix^i,$$

where $a_i\in\mathbb{K}$, $i=0,1,\ldots,k$ 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 $f:\mathbb{K}\to\mathbb{K}$ by substituting the argument c for the variable x and evaluating the resulting expression, i.e.

$$f(c)=a_0+a_1c+\cdots+a_kc^k.$$

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 $\mathbb{K}=\{a_0,a_1,\ldots,a_k\}$, the polynomial $f(x)=(x-a_0)(x-a_1)\ldots(x-a_k)$ defines the constant-zero mapping.

Polynomials $f(x)=\sum_i a_ix^i$ and $g(x)=\sum_i b_ix^i$ 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):

$$(f+g)(x)=(a_0+b_0)+(a_1+b_1)x+\cdots+(a_k+b_k)x^k,$$
$$(f\cdot g)(x)=(a_0b_0)+(a_0b_1+a_1b_0)x+\ldots$$
$$+(a_0b_r+a_1b_{r-1}+a_rb_0)x^r+\cdots+a_kb_\ell x^{k+\ell},$$

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.[5]

This definition corresponds to the addition and multiplication of the function values of $f,g:\mathbb{K}\to\mathbb{K}$, 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!

---

[5]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 atributed to have order $-\infty$.

polynomial, which means that $Q(x)$ divides $P(x)$. However, $P(x)$ is irreducible (by Eisenstein's criterion for 2), so its only 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 $P(x)$. □

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 $x^2 + 2x + 2$, $x^2 + x + 2$, $x^2 + 1$. □

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

$$x^4 + x^3 + x + 2 = (x^2 + ax + b)(x^2 + cx + d) =$$
$$= x^4 + (a + c)x^3 + (ac + b + d)x^2 +$$
$$+ (ad + bc)x + bd.$$

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 ℓ 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 $\mathbb{K}$. □

12.2.4. 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 $x_1, \ldots, x_r$ 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*

$$x_1^{k_1} \cdots x_r^{k_r},$$

and their formal linear combinations with coefficients $a_{k_1 \cdots k_r} \in \mathbb{K}$.

However, it is simpler, both formally and technically, to define them inductively by

$$\mathbb{K}[x_1, \ldots, x_r] := \big(\mathbb{K}[x_1, \ldots, x_{r-1}]\big)[x_r].$$

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 $x_1, \ldots, x_r$ can be viewed, even with this definition, as expressions created from the variables $x_1, \ldots, x_n$ 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

$$f = a_n(x)y^n + a_{n-1}(x)y^{n-1} + \cdots + a_0(x) =$$
$$= (a_{mn}x^m + \cdots + a_{0n})y^n + \cdots + (b_{p0}x^p + \cdots + b_{00}) =$$
$$= c_{00} + c_{10}x + c_{01}y + c_{20}x^2 + c_{11}xy + c_{02}y^2 + \ldots.$$

To simplify the notation, we use the multi-index notation (as we did with real polynomials and partial derivates in infinitesimal analysis).

860

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,

$$x^4 + x^3 + x + 2 = (x^2 + 1)(x^2 + x + 2). \qquad \square$$

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.. $\qquad \square$

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 $\mathbb{Z}_7[x]$. $\qquad \square$

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

---

A *multi-index* α of length r is an r-tuple of non-negative integers $(\alpha_1, \ldots, \alpha_r)$. The integer $|\alpha| = \alpha_1 + \cdots + \alpha_r$ is called the *size* of the multi-index α.

Monomials are written shortly as x^α instead of $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_r^{\alpha_r}$. Polynomials in r variables can be symbolically expressed in a similar way as univariate polynomials:

$$f = \sum_{|\alpha| \le n} a_\alpha x^\alpha, \ g = \sum_{|\beta| \le m} b_\beta x^\beta \in \mathbb{K}[x_1, \ldots, x_r].$$

f is said to have total degree n if at least one coefficient with multi-indices α 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:

$$f + g = \sum_{|\alpha| \le \max(m,n)} (a_\alpha + b_\alpha) x^\alpha,$$

$$fg = \sum_{|\gamma| = 0}^{m+n} \left(\sum_{\alpha + \beta = \gamma} (a_\alpha b_\beta) x^\gamma \right),$$

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

$$fg(x) = \sum_{i=0}^{k+\ell} \left(\sum_{p+q=i} a_p b_q x^i \right).$$

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 $\mathbb{K}[x_1, \ldots, x_{r-1}]$, and calculate the sum

$$f = a_k(x_1, \ldots, x_{r-1}) x_r^k + \cdots + a_0(x_1, \ldots, x_{r-1})$$

$$= \left(\sum_\alpha a_{k,\alpha} x^\alpha \right) x_r^k + \ldots,$$

$$g = b_l(x_1, \ldots, x_{r-1}) x_r^l + \cdots + b_0(x_1, \ldots, x_{r-1})$$

$$= \left(\sum_\beta b_{l,\beta} x^\beta \right) x_r^k + \ldots,$$

$$f + g = \big(a_0(x_1, \ldots, x_{r-1}) + b_0(x_1, \ldots, x_{r-1}) \big)$$

$$+ \big(a_1(x_1, \ldots, x_{r-1}) + b_1(x_1, \ldots, x_{r-1}) \big) x_r + \cdots$$

easy: $x^2 + 1 = (x+i)(x-1)$. 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

$$p(x) = (x+i)^2(x-i)^2\left(x - \sqrt{3}\right)\left(x + \sqrt{3}\right),$$

while in $\mathbb{R}[x]$, we have

$$p(x) = (x^2 + 1)^2 \left(x - \sqrt{3}\right)\left(x + \sqrt{3}\right).$$

In $\mathbb{Z}_5[x]$, the polynomial $x^2 + 1$ has roots ± 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 $\mathbb{Z}[x]$.

$$p(x) = (x^2 + 1)^2(x^2 - 3). \qquad \square$$

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., $\alpha = -\frac{1}{2} \pm \frac{1}{2}i\sqrt{7}$, then $y = \frac{1}{\sqrt{1+\alpha}}$. If $x = \beta$ is a root of $x^2 + 1$, i. e., $\beta = \pm i$, then $y = -\frac{1}{1+\beta}$ $\qquad \square$

12.C.12. Factorize the following polynomial to irreducible polynomials in $\mathbb{R}[x]$ and in $\mathbb{C}[x]$:

$$4x^5 - 8x^4 + 9x^3 - 7x^2 + 3x - 1.$$

$\bigcirc$

12.C.13. Factorize the following polynomial to irreducible polynomials in $\mathbb{R}[x]$ and in $\mathbb{C}[x]$:

$$x^5 + 3x^4 + 7x^3 + 9x^2 + 8x + 4.$$

$\bigcirc$

12.C.14. Factorize $x^4 - 4x^3 + 10x^2 - 12x + 9$ to irreducible polynomials in $\mathbb{R}[x]$ and in $\mathbb{C}[x]$:

$\bigcirc$

12.C.15. Decide whether the following polynomial over $\mathbb{Z}_3$ is irreducible; if not, factorize it to irreducible polynomials:

$$x^5 + x^2 + 2x + 1.$$

$\bigcirc$

$$= \left(\sum_\gamma (a_{k,\gamma} + b_{k,\gamma})(x_1, \ldots, x_{r-1})^\gamma\right) x_r^k + \ldots$$

$$+ \left(\sum_\gamma (a_{0,\gamma} + b_{0,\gamma})(x_1, \ldots, x_{r-1})^\gamma\right)$$

$$= \sum_{(\gamma, j)} (a_{j,\gamma} + b_{j,\gamma})(x_1, \ldots, x_{r-1})^\gamma x_r^j.$$

The proof for multiplication is similar (do it yourselves!). $\quad \square$

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 $\mathbb{K}[x_1, \ldots, x_r]$ is also an integral domain.*

PROOF. Proceed by induction on the number r of variables.[6] 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. $\quad \square$

Let us notice that each polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$ represents a mapping $\mathbb{K} \times \ldots \times \mathbb{K} \to \mathbb{K}$ symmetric in its n arguments.

12.2.5. 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).

<div align="center">DIVISIBILITY IN RINGS</div>

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 $0 = a \cdot (c - c')$

---

[6]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.

12.C.16. Decide whether the following polynomial over $\mathbb{Z}_3$ is irreducible; if not, factorize it to irreducible polynomials:

$$x^4 + 2x^3 + 2.$$

$\bigcirc$

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:

$$x^2 \pm 2,\, x^2 \pm x + 2,\, x^2 \pm 2x - 2,\, x^2 - x \pm 1,\, x^2 \pm 2x - 1.$$

$\square$

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. $\square$

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 $f = f_1 \ldots f_k$, then $\mathcal{V}(f) = \mathcal{V}(f_1) \cup \cdots \cup \mathcal{V}(f_k)$.

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. $\square$

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. $\square$

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) $\mathfrak{V}(x^3 + x^2 - y^2)$.

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

and being in an integral domain, $a \neq 0$ implies $c = c'$. The reader should check that the claim is wrong e.g. in $\mathbb{Z}_4$.

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 $a_1, \ldots, a_r \in \mathbb{K}$ such that $a = a_1 \cdot a_2 \ldots a_r$. This is called *factorization* of a.
- If there are two factorizations of a into irreducible non-unit elements $a = a_1 a_2 \ldots a_r = b_1 b_2 \ldots b_s$, 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, $j = 1, \ldots, r$.

$\mathbb{Z}$ 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

$$a_0 + \sum_{i=1}^{k} a_i \left(x^{m_i}\right)^{\frac{1}{2^{n_i}}},$$

where $a_0, \ldots, a_k \in \mathbb{Z}$, $m_i, n \in \mathbb{Z}_{>0}$. 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 ± 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 $\mathbb{K}$.

12.2.6. Euclidean division and roots of polynomials. The fundamental tool for the discussion of divisibility, common divisors, etc. in the ring of integers $\mathbb{Z}$ is the procedure of division with remainder, and the Euclidean algorithm for the greatest common divisor. These procedures can be generalized.

Consider univariate polynomials $a_k x^k + \cdots + a_0$ 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, $g \neq 0$.*
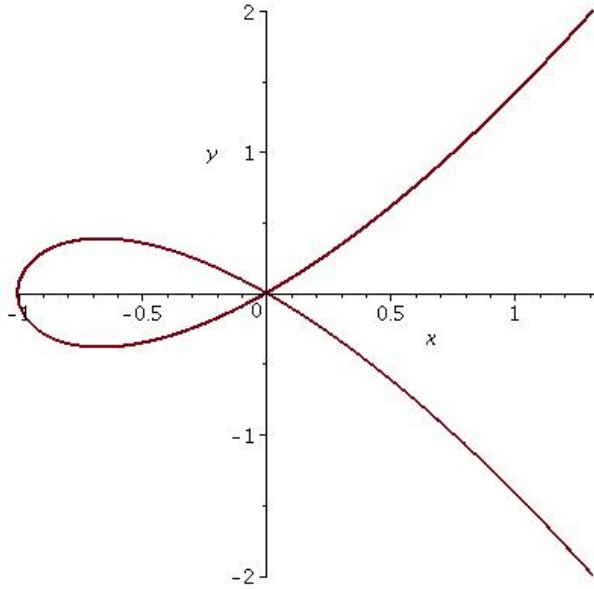
FIGURE 1. $\mathfrak{V}(x^3 + x^2 - y^2)$
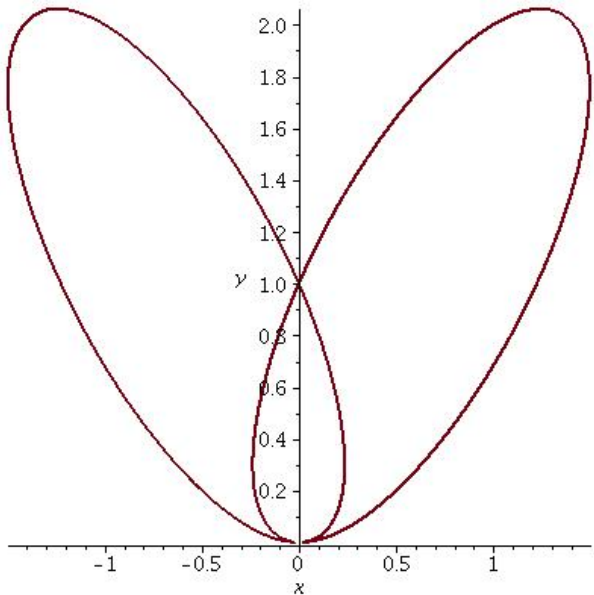
FIGURE 2. $\mathfrak{V}(2x^4 - 3x^2y + y^2 - 2y^3 + y^4)$

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. $\qquad\square$

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 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

$$f = a_0 + \cdots + a_nx^n, \quad g = b_0 + \cdots + b_mx^m.$$

Either $b_mf - a_nx^{n-m}g = 0$ or $\deg(b_mf - a_nx^{n-m}g) < \deg f$. 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

$$a'\left(b_mf - a_nx^{n-m}g\right) = q'g + r'$$

and either $r' = 0$ or $\deg r' < \deg g$. This means that

$$a'b_mf = \left(q' + a'a_nx^{n-m}\right)g + r'.$$

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,

$$b_mf = \left(q' + a_nx^{n-m}\right)g + r'.$$

If $\mathbb{K}$ is a field, then this equation can be multiplied by b_m^{-1}.

Assume that there is another solution $f = q_1g + r_1$. 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 $r = r_1$. $\qquad\square$

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. In particular, polynomials over an infinite integral domain define the same mapping $\mathbb{K} \to \mathbb{K}$ if and only if they are the same polynomial.*

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).
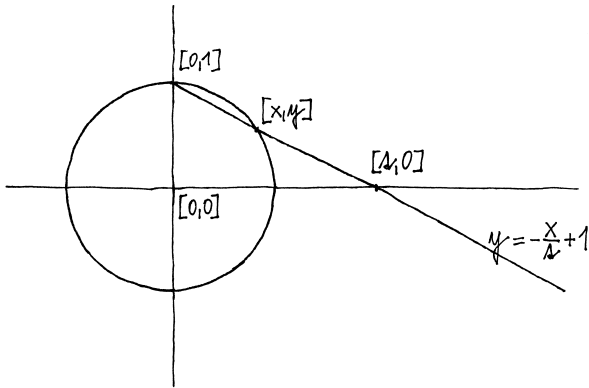
Solution. Substituting the equation of the line $y = \frac{x}{t} + 1$ into the equation of the circle, we get the equation

$$x^2 + \left(\frac{x^2}{t^2} - \frac{2x}{t} + \right) = 1,$$

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. $\quad\square$

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 obtain richer tools for description of their properties (imaginary tangent lines, etc.).

If two polynomials over an integral domain define the same mapping $\mathbb{K} \to \mathbb{K}$, 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.

12.2.7. Multiple roots and derivatives. We shall now work over infinite integral domains $\mathbb{K}$ and so we may identify the algebraic expressions for the polynomials with the mappings.

The differentiation of polynomials over real or complex numbers is an algebraic operation which make sense for all $\mathbb{K}$ and it still satisfies the Leibniz rule:

DERIVATIVE OF POLYNOMIALS

Consider polynomials $f(x) = a_0 + a_1 x + \cdots + a_n x^n$, $g(x) = b_0 + b_1 x \cdots + b_m x^m$ be polynomials of orders n and m over an commutative ring $\mathbb{K}$. The derivative $' : f(x) \mapsto f'(x) = a_1 + a_2 x + \cdots + n a_n x^{n-1}$ respects the addition of polynomials and their multiplication by the elements in $\mathbb{K}$. Moreover it satisfies the *Leibniz rule*

(1) $\qquad (f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$

While claim on the additive structure is obvious, let us check the Leibniz rule:

$$f(x) \cdot g(x) = \sum_{k=0}^{m+n} c_k x^k, \qquad c_k = \sum_{i+j=k} a_i b_j$$

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 $\mathbb{K}[x] \to \mathbb{K}[x]$ 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 are called *derivations*. For polynomial rings, we see inductively that the only derivative there is our operation $'$.

Differentiation can be used for discussing multiple roots of polynomials.

Consider a polynomial $f(x) \in \mathbb{K}[x]$ over infinite 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 12.2.6,

$$f(x) = (x - c)^k g(x),$$

with the unique polynomial $g, g(c) \neq 0$. Differentiating $f(x)$ and applying the Leibniz rule we obtain

$$f'(x) = k(x - c)^{k-1} g(x) + (x - c)^k g'(x)$$
$$= (x - c)^{k-1} \big(k g(x) + (x - c) g'(x) \big).$$

Clearly the polynomial $h(x) = k g(x) + (x - c) g'(x)$ does not admit c as root, i.e. $h(c) = k g(c) \neq 0$. Thus we have arrived at the following very useful claim

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^ε is given as a subset in $\mathbb{R}^4$ by a system of two real equations:

$$\mathrm{Re}(z_1^2 + z_2^2 - \varepsilon) = x_1^2 + x_2^2 - y_1^2 - y_2^2 - \varepsilon = 0,$$
$$\mathrm{Im}(z_1^2 + z_2^2 - \varepsilon) = 2(x_1y_1 + x_2y_2) = 0.$$

Thus, we can assume that X^ε 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 $\mathbb{R}^4 \to \mathbb{R}^3$. 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 φ_+ to V is invertible, and its inverse ψ_+ 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

$$\varphi_\varepsilon : X^\varepsilon \to V @> \varphi_+ >> \mathbb{R}^3 \setminus \{(0,0,0)\} \supseteq H^\varepsilon,$$

Proposition. *A polynomial $f(x)$ over an infinite integral domain $\mathbb{K}$ admits the root $c \in \mathbb{K}$ of multiplicity k if and inly if c is the root of $f'(x)$ of multiplicity $k - 1$.*

12.2.8. 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 (almost) 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 an 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

$$\varphi : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \frac{f(z)}{|f(z)|}.$$

Then φ 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 φ to the individual circles $K_r \subseteq \mathbb{C}$ with center at zero and radius $r \geq 0$
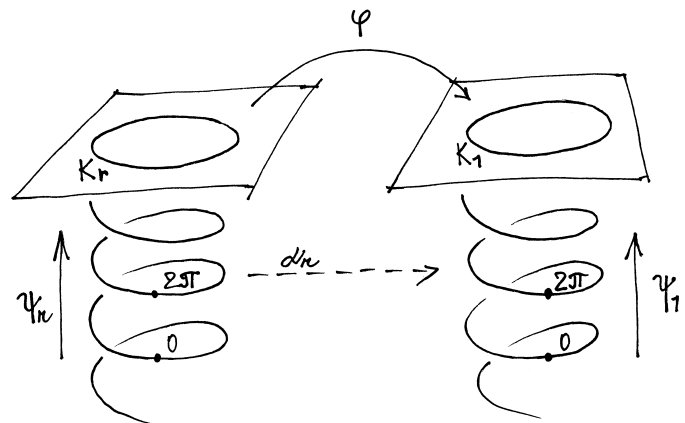
We can parameterize these circles by the mappings

$$\psi_r : \mathbb{R} \to K_r, \quad t \mapsto \psi_r(t) = r\,e^{it}.$$

For all r, the composition $\kappa : (0, \infty) \times \mathbb{R} \to K_1$, $\kappa(r, t) = \varphi \circ \psi_r(t)$, is continuous in both t and r. Thus, for each r, there exists a mapping $\alpha_r : \mathbb{R} \to \mathbb{R}$ which

and for every $\varepsilon > 0$, we get a bijection $\varphi_\varepsilon : X^\varepsilon \to H^\varepsilon$. The real part of this variety is the "thinnest circle" on the one-part rotational hyperboloid H^ε; see the picture.
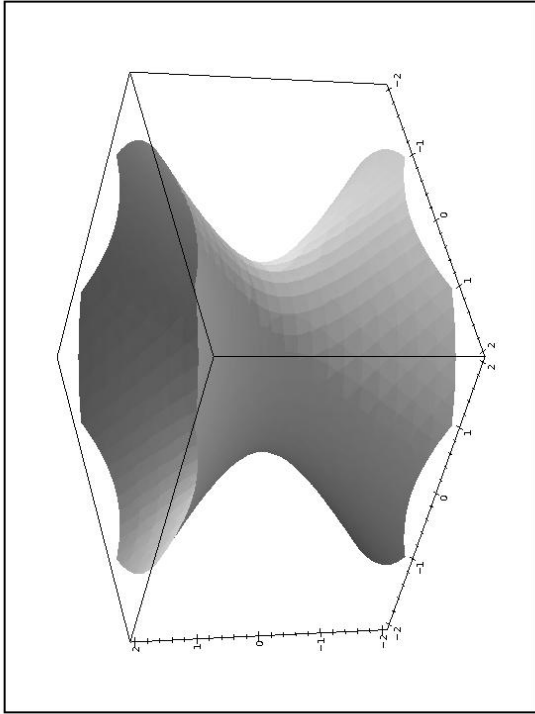
For $\varepsilon < 0$, we can repeat the above reasoning, merely interchanging x and y and the signs in the definition of φ_+:

$$\varphi_- : (x_1, x_2, y_1, y_2) \mapsto \left(-y_1, -y_2, \frac{-y_1 x_2 + y_2 x_1}{\sqrt{y_1^2 + y_2^2}} \right),$$

which changes the inversion ψ_-

$$\psi_+ : (u, v, w) \mapsto \left(-\frac{vw}{\sqrt{u^2 + v^2}}, \frac{uw}{\sqrt{u^2 + v^2}}, -u, -v \right).$$

Now, H^ε 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 $\mathbb{C}^2$.

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 α_r continuously depends on r. Altogether, there is a continuous mapping

$$\alpha : \mathbb{R} \times (0, \infty) \to \mathbb{R}, \quad (t, r) \mapsto \alpha_r(t).$$

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 α 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 $f = a_0 + \cdots + a_d z^d$ and $a_d \neq 0$, then for small values of r, α_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 $z = r(\cos\theta + i\sin\theta)$:

$$z^d = r^d(\cos d\theta + i\sin d\theta) = r^d\, e^{id\theta},$$

$$\frac{z^d}{|z^d|} = 1(\cos d\theta + i\sin d\theta) = e^{id\theta}.$$

In this case, the mapping φ 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 $f = a_0 + \cdots + a_d z^d$ with no root. Then $a_0 \neq 0$ ($a_0 = 0$, implies that 0 is a root). For $z \neq 0$,

$$\frac{f(z)}{a_d z^d} = 1 + \frac{1}{a_d}(a_0 z^{-d} + \cdots + a_{d-1} z^{-1}).$$

Hence, $\lim_{|z| \to \infty} \frac{f(z)}{a_d z^d} = 1$. Knowing this, calculate

$$\lim_{|z| \to \infty} \left| \frac{f(z)}{|f(z)|} - \frac{a_d z^d}{|a_d z^d|} \right| =$$

$$= \lim_{|z| \to \infty} \left| \frac{f(z)}{a_d z^d} \frac{a_d z^d}{|a_d z^d|} \frac{|a_d z^d|}{|f(z)|} - \frac{a_d z^d}{|a_d z^d|} \right| = 0.$$

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, $\lim_{|z| \to 0} \frac{f(z)}{a_0} = 1$. In addition, $\frac{f(z)}{|f(z)|} = \frac{f(z)}{a_0} \frac{a_0}{|a_0|} \frac{|a_0|}{|f(z)|}$. Hence, $\lim_{|z| \to 0} \frac{f(z)}{|f(z)|} = \lim_{|z| \to 0} \frac{a_0}{|a_0|}$, i.e. $n_r = 0$ for small values of r. Altogether, the degree of the polynomial is $d = 0$. $\qquad\square$

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.

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:
$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$D = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, E = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, F = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \text{ The table}$$

12.2.9. The greatest common divisor. Consider a polynomial ring $\mathbb{K}[x]$ over an integral domain $\mathbb{K}$. A polynom $h \in \mathbb{K}[x]$ 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 12.2.6):

$$f = q_1 g + r_1,$$
$$g = q_2 r_1 + r_2,$$
$$r_1 = q_3 r_2 + r_3,$$
$$\vdots$$
$$r_{p-1} = q_{p+1} r_p + 0.$$

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:

$$h = r_p = r_{p-2} - q_p r_{p-1} =$$
$$= r_{p-2} - q_p(r_{p-3} - q_{p-1}r_{p-2}) =$$
$$= -q_p r_{p-3} + (1 + q_{p-1}q_p)r_{p-2} =$$
$$= -q_p r_{p-3} + (1 + q_p q_{p-1})(r_{p-4} - q_{p-2}r_{p-3}) =$$
$$\vdots$$
$$= Af + Bg. \qquad \square$$

12.2.10. 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.

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. □

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) $(\mathbb{Z}_9, \cdot)$.

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 $f = f_1 \ldots f_r$ is obtained, where the polynomials $f_1, \ldots, f_r$ are irreducible.

It follows from the additional assumption that every irreducible polynomial h which divides f also divides one of $f_1, \ldots, f_r$. Therefore, for every other factorization $f = f_1' f_2' \ldots f_s'$, 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. □

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, $\mathbb{K}[x_1, \ldots, x_r]$ 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.

12.2.11. 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.

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

$$\frac{a}{b} = \frac{a'}{b'} \quad \Leftrightarrow \quad ab' = a'b.$$

Addition and multiplication is defined in terms of representatives:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd},$$
$$\frac{a}{b}\frac{c}{d} = \frac{ac}{bd}.$$

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) $a \heartsuit b = a + (-1)^a b$.

○

12.E.6. In how many ways can we fill the following table so that $(\{a, b, c\}, \divideontimes)$ would be a

  i) groupoid   iv) a monoid
  ii) commutative groupoid   v) a group
  iii) a groupoid with a neutral
       element

$\divideontimes$	a	b	c
a	c	b	a
b			b
c			

Solution.

  i) 3^5 iv) 1
  ii) 9   v) 0
  iii) 9

□

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 $\mathbb{K}[x_1, \ldots, x_r]$ is called the *field of rational functions* (of r variables) and denoted $\mathbb{K}(x_1, \ldots, x_r)$. In software systems like 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 $\mathbb{K} = \mathbb{Q}$.

12.2.12. 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 $d = d_1 \ldots d_r$, $b = b_1 \ldots b_s$, $c = c_1 \ldots c_q$ be the factorizations to irreducible factors. This means that

$$ad_1 \ldots d_r = b_1 \ldots b_s c_1 \ldots c_q.$$

Since ad factors in a unique way, it follows that $a = eb_j$ or $a = ec_i$ for a suitable unit e.

(2) Let $f = b_0 + b_1 x + \cdots + b_n x^n$. Since $a|f$, there must exist a polynomial $g = c_0 + c_1 x + \ldots c_k x^k$ such that $f = ag$. Hence it immediately follows that $k = n$, $ac_0 = b_0, \ldots, ac_n = b_n$.

(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 $b_0 c_{i+j} + b_1 c_{i+j-1} + \cdots + b_{i+j} c_0$. By choice, a divides all of $b_0 c_{i+j}, \ldots, b_{i-1} c_{j+1}, b_{i+1} c_{j-1}, \ldots, b_{i+j} c_0$. At the same time, it does not divide $b_i c_j$. Therefore, it cannot divide the coefficient. □

12.2.13. 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

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

$$3^{3\cdot3} = 19683$$

groupoids. □

12.E.8. Decide whether the set $G = (\mathbb{R} \smallsetminus \{0\} \times \mathbb{R})$ together with the operation $\triangle$ defined by $(x, y)\triangle(u, v) = (xu, xv+y)$ for all $(x, y), (u, v) \in G$ is a groupoid, semigroup, monoid, group, and whether $\triangle$ is commutative. ○

12.E.9. Let Ω be a set with the multiplication operation defined by $ab = a$ for all a, b in Ω. 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) $\pi^4 = (1, 2, 3, 4, 5, 6, 7)$
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) $\begin{pmatrix} 1 & 2 & 3 & \dots & n & n+1 & n+2 & \dots & 2n \\ 2 & 4 & 6 & \dots & 2n & 1 & 3 & \dots & 2n-1 \end{pmatrix}$.

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. □

12.2.14. 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 12.2.10.

3. Groups

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.

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 inverses 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 σ in the group $(\mathbb{S}_7, \circ)$, the inverse of σ and compute σ^{2013}. Show that σ does not commute with the transposition $\tau = (2, 3)$.

Solution. $\sigma = (1, 3, 5) \circ (2, 6) \circ (4, 7)$. Therefore, the order of σ 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 σ^{-1} and σ^{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 σ is the least common multiple of 2 and 5, which is 10, i. e., $\sigma^{10} = 1$. Then,

$$\sigma^{2013} = (\sigma^{10})^{201} \circ \sigma^3 = \sigma^3 = (1, 2, 4, 5, 6) \circ (3, 7)$$

.

12.3.1. Examples and concepts. Let A be a set. A *binary operation* on A is defined to be any mapping $A \times A \to A$. 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 from 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 $b - a$.

NEUTRAL ELEMENTS, INVERSES, AND GROUPS[7]

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 grupoid $(A, \cdot)$ 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.

---

[7]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.

(b) For the sake of simplicity, we will write only k for the residue class $[k]_{11}$, $k \in \mathbb{Z}$. Then,

$$4^5 \equiv 1 \pmod{1}1 \Rightarrow \sigma^{-1} = 4^4 \equiv 3 \pmod{1}1$$
$$\sigma^{2013} = 4^{2013} \equiv 4^3 \equiv 9 \pmod{1}1.$$

$\square$

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. $\square$

12.F.8. Prove that there exists no non-commutative group of order 4.

Solution. By Lagrange's theorem (see 12.3.10), 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 $\mathbb{Z}_2 \times \mathbb{Z}_2$.

$\square$

12.F.9. Show that there exists no non-commutative group of order 5.

Solution. By Lagrange's theorem (see 12.3.10), the non-trivial elements of a 5-element group are of order 5, so the group must be cyclic, and thus commutative. $\square$

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

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 ± 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.

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 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 $\mathrm{Hom}(V, V)$ 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.

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.

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

$$ab = ab((ba)(ba)) = a(bb)aba = (aa)ba = ba.$$

$\square$

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 12.3.10), 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 12.3.10, 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

12.3.2. 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 $\alpha : M \to M$ is required to have an inverse α^{-1}, then α must be a bijection. Composition of two bijections yields again a bijection; hence the set Σ_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.[8]

The name of the group Σ_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 81 for a few 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 Σ_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!)

$$a = (1, 2, 3), \ b = (2, 3, 1), \ c = (3, 1, 2),$$
$$d = (1, 3, 2), \ e = (3, 2, 1), \ f = (2, 1, 3),$$

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:

$$b^2 = c, \ b^3 = a, \ c^2 = b, \ c^3 = a.$$

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.

---

[8]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 Σ_m are as non-commutative and complex as possible.

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 $120°$), see also 12.3.3.

We have discussed all possibilities, so the proof is finished. □

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 12.3.8, every commutative group is a product of cyclic groups. By 12.3.10, 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 $\mathbb{Z}_2 \times \mathbb{Z}_4$.
- $\mathbb{Z}_{17}^{\times} = \{\pm 1, \pm 2, \ldots, \pm 8\}$. Thus, the quotient $\mathbb{Z}_{17}^{\times}/(\pm 1) = \{1, 2, \ldots, 8\}$ 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 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 Σ_m of finite m-element sets behave similarly. Each permutation σ 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 $\sigma^k(x)$, $k = 1, 2, \ldots$, 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 σ on M_x. If the elements of M_x are numbered as $(1, 2, \ldots, |M_x|)$ 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 σ 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 σ 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 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 Σ_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 82).

It follows that there is a well-defined mapping

$$\text{sgn} : \Sigma_m \to \mathbb{Z}_2 = \{\pm 1\},$$

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 ℓ 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 σ and τ.

- The complex roots of the polynomial $z^8 - 1$ are $e^{\frac{n\pi}{4}i}$, where $n = 1, 2, \ldots, 8$. Clearly, these form a cyclic group of order 8, isomorphic to $\mathbb{Z}_8$. $\quad\square$

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 $(ab)^2 = a^2 b^2 = e$ (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. $\quad\square$

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) $H = \left\{ \begin{pmatrix} 0 & a \\ b & c \end{pmatrix} \in G \mid a, b, c \in \mathbb{R} \right\}$,

viii) $H = \left\{ \begin{pmatrix} 1 & a \\ b & c \end{pmatrix} \in G \mid a, b, c \in \mathbb{R} \right\}$.

$\bigcirc$

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 $(\mathbb{R}, +)$.

$\bigcirc$

12.F.16. Find all positive integers $m \neq 5$ such that the group $\mathbb{Z}_m^\times$ is isomorphic to $\mathbb{Z}_5^\times$. $\quad\bigcirc$

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,

The last proposition suggests that the mapping sgn transforms permutation composition $\sigma \circ \tau$ to the product $\operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$ in the commutative group $\mathbb{Z}_2$.

(SEMI)GROUP HOMOMORPHISMS

In general, a mapping $f : G_1 \to G_2$ 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 $\operatorname{sgn} : \Sigma_m \to \mathbb{Z}_2$.

In a moment, we shall see that the group inversions and units are also preserved by homorphisms. Before disussing the theory, let us look at more examples of groups.

12.3.3. 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 $\mathrm{Mat}_2(\mathbb{R})$ defines a linear mapping $\mathbb{R}^2 \to \mathbb{R}^2$ 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.

In fact, it is possible to prove that every mapping of the plane into itself which preserves distances is affine such a Euclidean transformation.[9]

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_φ around the origin by any angle φ, and reflection F_ℓ with respect to any line that goes through the origin (also note that the central inversion is the same as the rotation by π).

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.

---

[9]If a mapping $F : \mathbb{R}^2 \to \mathbb{R}^2$ 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 $F = (f(x, y), g(x, y)) : \mathbb{R}^2 \to \mathbb{R}^2$ 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.

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. $\qquad\square$

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

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix}. \qquad\square$$

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 $C_G(C_G(X)) = \{y \in G \mid xy = yx, \forall x \in C_G(X)\}$, 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 $C_G(X) \supseteq C_G(C_G(C_G(X)))$. $\qquad\square$

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.

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).
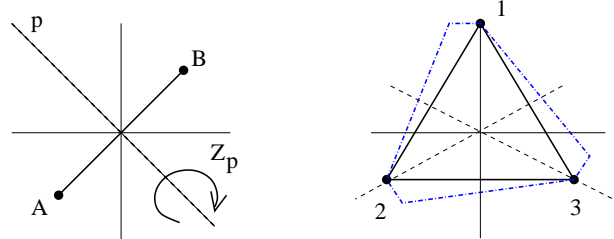
In the case of the line segment, the situation is very simple – it is clear that the only non-trivial symmetries are rotation by π around the center of the segment, reflection F_H with respect to the axis of the segment, and reflection F_V with respect to the line itself. All these symmetries are self-inverse. Hence the group of symmetries has four elements. Its table looks as follows:

$\cdot$	R_0	R_π	F_H	F_V
R_0	R_0	R_π	F_H	F_V
R_π	R_π	R_0	F_V	F_H
F_H	F_H	F_V	R_0	R_π
F_V	F_V	F_H	R_π	R_0

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 Σ_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

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. □

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 $|G| = 2$. □

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 $(ab)^{rs} = a^{rs}b^{rs} = (a^r)^s(b^s)^r = e^se^r = e$, 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

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 π 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 $k > 1$.*

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 2π, then iterating this rotation would lead to a dense subset of images on the corresponding circle.[10] 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

---

[10]The argument is subtle but straightforward: if there was an interval of length ε 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 ε. Thus there can be only finitely many of them and this contradicts the irrationality of the angle.

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 $(\mathbb{Q}, +) \to (\mathbb{Q}^+, \cdot)$.

Solution. There is only one homomorphism, the trivial one. For the sake of contradiction, assume that there exists a non-trivial homomorphism φ, 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. 1.G.1). $\square$

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:

$$\begin{pmatrix} a & 0 \\ b & a^{-1} \end{pmatrix} \begin{pmatrix} a_1 & 0 \\ b_1 & a_1^{-1} \end{pmatrix} = \begin{pmatrix} aa_1 & 0 \\ ba_1 + a^{-1}b_1 & a^{-1}a_1^{-1} \end{pmatrix}.$$

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 φ 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 φ_0 (there are only finite many of them there). But then it is impossible to allow a rotation R_φ where φ is not a multiple of φ_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 φ_0). This subgroup coincides with one of C_ℓ. Next, adding one reflection produces exactly one different reflection for each nontrivial element in C_ℓ, as seen above. $\square$

12.3.4. 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).

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_π, 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 π.

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:

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 2×2 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.

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. Nence, index $H \in G$ equals 2 and, therefore, $|H| = \frac{|G|}{2} = 24$.

$\square$

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 $n_1 h_1$ and $n_2 h_2$ are two elements in NH Then

i) $n_1 h_1 n_2 h_2 = n_1 h_1 n_2 (h_1)^{-1} h_1 h_2 = n_1 n_2' h_1 h_2 = n' h' \in NH$.

ii) $(nh)^{-1} = h^{-1} n^{-1} = h^{-1} n^{-1} h^{-1} h = n' h$.

$\square$

12.F.28. Suppose that N and M are two normal subgroups of G and $N \ cup M = \{e\}$. Show that $nm = mn$ for any $n \in N$ and $m \in M$.
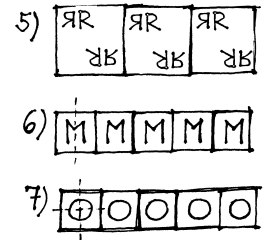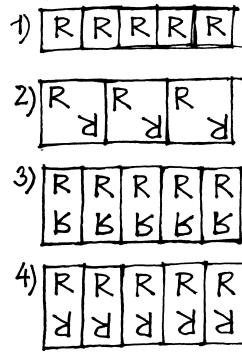
It is even more complicated with symmetries of tilings which cover the entire plane. There is insufficient space here to consider further details. It can be shown that there are 17 such groups of symmetries, known as the two-dimensional crystallographic groups.

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.

12.3.5. Group homomorphisms. Recall that a mapping $f : G \to H$ 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 $f : G \to H$ 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.

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 $mnm^{-1}n^{-1} in\, M \cup N$. So, $mnm^{-1}n^{-1} = e$. Thus, $mn = nm$. $\qquad\square$

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 $gng^{-1} \in N$ and $gmg^{-1}1 \in M$. Let $w \in N \cup M$. Since $w \in N$ we have $gwg^{-1} \in N$ and since $w \in M$ we have $gwg^{-1} \in M$. Hence $gwg^{-1} \in N \cup M$ for any $g \in G$ i.e. $N \cup M$ is a normal subgroup of G. $\qquad\square$

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, $xH = H = Hx$ 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 $xH = (ah1)H = aH = Hb = H(h_2b) = Hx$, i.e. $xH = Hx$ for any $x \in G$, so $G^{-1}HG = H$. $\qquad\square$

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 $gmg^{-1}1 \in M$. Let $w \in N \cup M$. Since $w \in N$ we have $gwg^{-1} \in N$ and since $w \in M$ we have $gwg^{-1} \in M$. Hence $gwg^{-1} \in N \cup M$ for any $g \in G$ i.e. $N \cup M$ is a normal subgroup in G. $\qquad\square$

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 $gng^{-1} \in N$. $\qquad\square$

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. $\qquad\square$

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 $f : G \to H$ with a trivial kernel is an isomorphism onto the image $f(G)$.

12.3.6. 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 $\exp : \mathbb{R} \to \mathbb{R}_+$ 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 $\exp : \mathbb{C} \to \mathbb{C}\setminus\{0\}$ 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 $\det : G \to \mathbb{K} \setminus \{0\}$ is a group homomorphism.

12.3.7. Group product. Given any two groups, a more complicated group can be constructed using the following construction:

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 $xN^k = x^k N$ for any $k \in Z$, then $N = (xN)^{|G/N|} = x^{|G/N|}N$, which is the identity element in $G\mathbb{N}$. This means $x^{|G/N|} \in N$ and so $x = (x^{|G:N|})^m \in N$. $\qquad \square$

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 $ah2a^{-1} = h_1^{-1}h_3 = h' \in H$ for any $h_2 \in H$ and $a \in G$. So, H is normal in G. $\qquad \square$

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 $yx = xy$. $\qquad \square$

12.F.36. Let G be a group of 2×2 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 $H = \left\{ \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} | x \in \mathbb{Z} \right\} \in G$ 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, $g^{-1}Hg = \left\{ \begin{pmatrix} 1 & 0 \\ 2x & 1 \end{pmatrix} | x \in \mathbb{Z} \right\}$ does not contain $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in H$. $\qquad \square$

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

$$\ker p_G = \{(e_G,b);\ b \in H\} \simeq H,$$
$$\ker p_H = \{(a,e_H);a \in G\} \simeq G.$$

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 ± 1, while $\mathbb{Z}_3$ corresponds to the equilateral triangle, one of whose vertices is 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):

$$[0]_6 \mapsto ([0]_2, [0]_3),$$
$$[1]_6 \mapsto ([1]_2, [2]_3),$$
$$[2]_6 \mapsto ([0]_2, [1]_3),$$
$$[3]_6 \mapsto ([1]_2, [0]_3),$$
$$[4]_6 \mapsto ([0]_2, [2]_3),$$
$$[5]_6 \mapsto ([1]_2, [1]_3).$$

Similar constructions are available for finite commutative groups in complete generality.

12.3.8. 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.

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 Ω 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 $|G| \equiv \left|\{g \in G | g^2 = e\}\right| \mod 2$. One solution to $g^2 = e$ is e itself. If it were the only solution, then $|G| \equiv 1 \mod 2$, 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 $G = \{g_1, g_2, \ldots, g_n\}$ 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 $m_i, i = 1, 2, \ldots, n$, so m is not divisible by p and $g^{m_i} = e$ for all i. Because G is abelian, the function $f : (\mathbb{Z}/m)^n \to G$ given by

$$f(a_1, ..., a_n) = g_1^{a_1} \ldots g_n^{a_n}.$$

is a homomorphism $f(a1, ..., an) f(b1, ..., bn) = f(a_1 + b_1, \ldots, a_n + b_n)$. as, by commutativity of G,

$$g^{a_1} \ldots g_n^{a_n} g_1^{b_1} \ldots g_n^{b_n} = g_1^{a_1+b_1} \ldots g_n^{a_n+b_n}$$

's. This homomorphism is surjective (each element of $g_i \in G$ is $F(a_1, \ldots, a_n)$ for $a_i = 1$ and other $a_j = 0, j \neq i$. g_i , and if ai = 1 and other aj 's are 0 then f(a1, . . . , an) = gi) and the elements where f takes particular value g_i is a coset of ker f, so $|G| = $ 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. $\square$

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,

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 $n = p_1^{k_1} \cdots p_r^{k_r}$ is the prime factorization of n, then the group C_n is isomorphic to the product

$$C_n = C_{p_1^{k_1}} \times \cdots \times C_{p_r^{k_r}}.$$

INCOMPLETE PROOF. We are going to prove only the second claim now, and we return to the first claim later, see 12.3.12, 12.3.13.

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 $f : C_n \to C_p \times C_q$ 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 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.

12.3.9. 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 $a^{-1} \cdot b = (b^{-1} \cdot a)^{-1} = h^{-1} \in H$,

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 $p|[G : H]$ for every proper subgroup H. Denote conjugacy class of an element $a \in G$ by $C(a) = \{b \in G | b = g^{-1}ag \text{ for some } g \in G\}$. Let the conjugacy classes in G with size greater than 1 be represented by $g_1, g_2, \ldots, g_k$. 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 $\forall j, \quad |G| = |Z(g_i)|[G : Z(g_i)]$, it yields $p|[G : Z(g_i)]$. 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 12.3.10, 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 $[a]^k$, $k = 0, \ldots, 6$, i. e., $[a^k]$. Then, each element of G

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 $a \in G$, $h \in H$

$$a \cdot h \cdot a^{-1} \in H.$$

(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:

12.3.10. 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.

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. $\bigcirc$

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 $p : H \to \mathbb{Z}_2$ which maps each permutation to its parity (0 for even and 1 for odd). Then, $p^{-1}(0) = \mathrm{Ker}(p)$ is a normal subgroup of H: let $h \in \mathrm{Ker}(p)$, 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 $ghg^{-1} \in \mathrm{Ker}(p)$, i. e., $\mathrm{Ker}(p)$ is normal. Since Z_2 has only two elements, it follows that $H/\mathrm{Ker}(p)$ 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 12.3.10). 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

Each element $a \in G$ generates the cyclic subgroup $\{a, a^2, \ldots, a^k = e\}$, 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^ℓ for $\ell = 1, \ldots, n$. $\square$

12.3.11. 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:

$$(a \cdot h \cdot b \cdot h') \cdot H = ((a \cdot b) \cdot (b^{-1} \cdot h \cdot b) \cdot h') \cdot H = (a \cdot b) \cdot H.$$

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 straighforward. 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

$$p : G \to G/H, \quad a \mapsto a \cdot H$$

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 $f : G \to K$ with kernel $H = \ker f$, there is a well-defined homomorphism

$$\tilde{f} : G/H \to K, \, \tilde{f}(a \cdot H) = f(a),$$

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.

double transpositions (compositions of two disjoint transpositions). Geometrically, double transpositions correspond to rotation by $180°$ around the axis that goes through the centers of opposite edges). Thus, we get nine subgroups: $\{\mathrm{id}, (a,b)\}$, $\{\mathrm{id}, (a,c)\}$, $\{\mathrm{id}, (a,d)\}$, $\{\mathrm{id}, (b,c)\}$, $\{\mathrm{id}, (b,d)\}$, $\{\mathrm{id}, (c,d)\}$, $\{\mathrm{id}, (a,b) \circ (c,d)\}$, $\{\mathrm{id}, (a,c) \circ (b,d)\}$, $\{\mathrm{id}, (a,d) \circ (b,c)\}$.

(ii) $|H| = 3$. By Lagrange's theorem, such a subgroup must be cyclic, i. e., it must be of the form $\{\mathrm{id}, p, p^2\}$, $p^3 = \mathrm{id}$. 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 subgroups: $\{\mathrm{id}, (a,b,c), (a,c,b)\}$, $\{\mathrm{id}, (a,c,d), (a,d,c)\}$, $\{\mathrm{id}, (a,b,d), (a,d,b)\}$, $\{\mathrm{id}, (b,c,d), (b,d,c)\}$. The cycles of length 3 correspond to rotation by $120°$ 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:

$$\{\mathrm{id}, (a,b,c,d), (a,c) \circ (b,d), (a,d,c,b)\},$$
$$\{\mathrm{id}, (a,c,b,d), (a,b) \circ (b,d), (a,d,b,c)\},$$
$$\{\mathrm{id}, (a,b,d,c), (a,d) \circ (bc), (a,c,d,b)\}.$$

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 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: $\{\mathrm{id}, (a,b), (a,b) \circ (c,d), (c,d)\}$, $\{\mathrm{id}, (a,c), (a,c) \circ (b,d), (b,d)\}$, $\{\mathrm{id}, (a,d), (a,d) \circ (b,c), (b,c)\}$ and $\{\mathrm{id}, (a,b) \circ (c,d), (a,c) \circ (b,d), (a,d) \circ (b,c)\}$.

There is a seemingly paradoxical example of a group homomorphism $\mathbb{C}^* \to \mathbb{C}^*$, 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

$$\tilde{f} : \mathbb{C}^*/\mathbb{Z}_k \to \mathbb{C}^*$$

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 12.3.10.

12.3.12. Exact sequences. A normal subgroup H of a group G yields the *short exact sequence of groups*

$$e \to H \to G \to G/H \to e,$$

where the arrows respectively correspond to the only homomorphism of the trivial group $\{e\}$ into the group H, the inclusion ι of the subgroup $H \subseteq G$, the projection ν 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.

If there exists a homomorphism $\sigma : G/H \to G$ such that $\nu \circ \sigma = \mathrm{id}_{G/H}$, it is said that the exact sequence *splits*.

Lemma. *Every split short exact sequence of commutative groups defines an isomorphism $G \to H \times G/H$.*

Proof. Define a mapping $f : H \times G/H \to G$ 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 12.3.8 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 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.

(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 $\{\mathrm{id}, (a, b) \circ (c, d), (a, c) \circ (b, d), (a, d) \circ (b, c)\}$, 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 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 $\mathbb{S}_4$. However, we could see in (iv) that there is no 6-element subgroup of $\mathbb{S}_4$ consisting only of even permutations. Thus, the only possibility is the alternating group $\mathbb{A}_4$ of all even permutations in $\mathbb{S}_4$. From the

12.3.13. 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.*[11]

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 ℓp for some integer ℓ. Hence the element g^ℓ has the required order p. $\square$

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 than 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^ℓ with q not divisible by p. Then g^{p^ℓ} 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.

---

[11]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.

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. □

Remark. In general, the group of symmetries of a solid with n vertices is a subgroup of the symmetric group $\mathbb{S}_n$.

12.F.45. Which subgroups of the group $\mathbb{S}_4$ are normal?

Solution. By definition, a subgroup $H \subseteq \mathbb{S}_4$ is normal iff it is closed under conjugation, i. e., $ghg^{-1} \in H$ for any $g \in \mathbb{S}_4$, $h \in H$. 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 $\mathbb{A}_4$ of all even permutations, and the entire group $\mathbb{S}_4$. □

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° around the axis that lies in the plane and goes through the centers of two opposite sides. The group is isomorphic to $\mathbb{S}_4 \times \mathbb{Z}_2$. □

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° around the axis that lies in the plane and goes through the centers of two opposite sides. Is this subgroup normal? ◯

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 $\mathbb{A}_4$

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. □

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 $\sigma : G \to G$, $\sigma(g) = g^p$ 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 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. □

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 $H \subset L \subset$

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

$$\{(1,2,4),(2,4,3),(1,3)\circ(2,4)\},$$
$$\{(1,4,2),(1,4,3),(1,4)\circ(2,3)\},$$
$$\{(2,3,4),(1,2)\circ(3,4),(1,3,4)\},$$
$$\{\mathrm{id},(1,2,3),(1,3,2)\}.$$

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.) □

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 $\mathbb{S}_4 \times S_2 \subseteq S_6$ (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

$$(1,2,3,4)^2 = (1,3)\circ(2,4),$$
$$(1,2,3,4)^3 = (4,3,2,1),$$
$$(1,2,3,4)^4 = \mathrm{id},$$
$$[(1,2)\circ(3,4)]^2 = \mathrm{id},$$
$$[(1,2)\circ(3,4)]\circ(1,2,3,4) = (2,4),$$
$$(1,2,3,4)\circ[(1,2)\circ(3,4)] = (1,3),$$
$$[(1,2)\circ(3,4)]\circ(4,3,2,1) = (1,3),$$
$$(4,3,2,1)\circ[(1,2)\circ(3,4)] = (2,4),$$

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 $G = C \times L$. □

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 $H_1 \times \ldots \times H_k$ and $H'_1 \times \ldots \times H'_\ell$ 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 $g_1, \ldots, g_t$ are generators of the entire G, then all elements of G are of the form $a_1 g_1 + \cdots + a_t g_t$ 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 $G = \mathbb{Z}^\ell \oplus \mathbb{Z}_{p_1} \times \ldots \times \mathbb{Z}_{p_k}$. This means there is always a finite number of completely independent generators of G and each of them generates a cyclic subgroup in G. (Compare this to the description of finite dimensional vector spaces via their basis, as discussed in chapter 2.)

12.3.14. 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 2π is the identity as a mapping.

Formally, this situation can be described as follows:

$$[(1,2) \circ (3,4)] \circ [(1,3) \circ (2,4)] = (1,4) \circ (2,3),$$
$$[(1,3) \circ (2,4)] \circ [(1,2) \circ (3,4)] = (1,4) \circ (2,3),$$
$$[(1,2) \circ (3,4)] \circ (4,2) = (1,2,3,4),$$
$$(1,3) \circ (4,2) = (1,3) \circ (2,4).$$

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 $H \subseteq \mathbb{S}_4$ has 8 elements (which is possible by Lagrange's theorem, since 8 divides 24).

$$H = \{\, \mathrm{id}, (1,2,3,4), (1,3) \circ (2,4), (4,3,2,1), (1,2) \circ (3,4),$$
$$(1,3), (2,4), (1,4) \circ (2,3)\,\}.$$

Altogether, the examined subgroup in $\mathbb{S}_6$ has 16 elements: for each $h \in H$, it contains (h, id) and $(h, (56))$. $\square$

12.F.50. Find the subgroup in $\mathbb{S}_4$ 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

$$[(1,2) \circ (3,4)]^2 = \mathrm{id},$$
$$(1,2,3)^2 = (3,2,1),$$
$$[(1,2) \circ (3,4)] \circ (1,2,3) = (2,4,3),$$
$$(1,2,3) \circ [(1,2) \circ (3,4)] = (1,3,4),$$
$$[(1,2) \circ (3,4)] \circ (3,2,1) = (3,1,4),$$
$$(3,2,1) \circ [(1,2) \circ (3,4)] = (2,3,4),$$

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. $\square$

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 (12.3.10), 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 *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 $S \to S$. Such a homomorphism can be viewed as a mapping $\varphi : G \times S \to S$ 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 φ (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 φ 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 $G_x \to G_y$.

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.

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 $G \times S \to S$ 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.

12.3.15. Theorem. *Let an action of a finite group G on a finite set S be given. Then:*

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 $BDB = E$, $CDC = E$, $FDF = E$ (whence it follows that $BEB = D, CEC = D$, $FEF = D$), we find out that this subgroup is normal. $\quad\square$

12.F.52. Find all subgroups of the group $(\mathbb{Z}_{10}, +)$.

Solution. The subgroups are isomorphic to $(\mathbb{Z}_d, +)$, where $d \mid 10$, i. e., $\{0\} \cong \mathbb{Z}_1$, $\{0, 5\} \cong \mathbb{Z}_2$, $\{0, 2, 4, 6, 8\} \cong \mathbb{Z}_5$, and $\mathbb{Z}_{10}$. $\quad\square$

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 $x^k \equiv 1 \mod 35$. 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 $k \cdot x \equiv 0 \mod 35$. This can be calculated simply as $k = \frac{35}{(35, x)}$. Therefore, the order of 2 and 4 is 35, while the order of 5 is 7. $\quad\square$

12.F.54. Find all finite subgroups of the group $(\mathbb{R}^*, \cdot)$[1]

Solution. If a given subgroup of the group $(\mathbb{R}^*, \cdot)$ contains an element a, $|a| \neq 1$, then the elements $a, a^2, a^3, \ldots$ 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: $\{1\}, \{-1, 1\}$. $\quad\square$

---

[1] 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 $\mathbb{Z}_n$.

(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 $G/G_x \to S_x$, $g \cdot G_x \mapsto g \cdot x$. 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 $x_1, \ldots, x_N$), 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. $\quad\square$

It is recommended to think out thoroughly how useful these propositions are for solving combinatorial problems, c.f. 12.G.1 through 12.G.4.

4. 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.

12.F.55. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism, and if so, find its kernel. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \to \mathbb{Z}_{12}$, $\varphi(([a]_4, [b]_3)) = [a - b]_{12}$,

ii) $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \to \mathbb{Z}_{12}$, $\varphi(([a]_4, [b]_3)) = [6a + 4b]_{12}$,

iii) $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \to \mathbb{Z}_{12}$, $\varphi(([a]_4, [b]_3)) = [0]_{12}$.

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 $\mathrm{Ker}(\varphi)$ is the set $\{([2]_4, [0]_3), ([0]_4, [0]_3)\}$.

iii) A homomorphism, neither injective, nor surjective. Its kernel is the entire group $\mathbb{Z}_4 \times \mathbb{Z}_3$. □

12.F.56. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism, and if so, find its kernel. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathbb{Z}_4 \to \mathbb{C}^*$, $\varphi([a]_4) = i^a$,

ii) $\varphi : \mathbb{Z}_5 \to \mathbb{C}^*$, $\varphi([a]_5) = i^a$,

iii) $\varphi : \mathbb{Z}_4 \to \mathbb{C}^*$, $\varphi([a]_4) = (-1)^a$,

iv) $\varphi : \mathbb{Z} \to \mathbb{C}^*$, $\varphi(a) = i^a$.

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 $\mathrm{Ker}(\varphi) = \{[0]_4\}$), 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 $\mathrm{Ker}(\varphi) = \{[0]_4, [2]_4\}$.

iv) Is a homomorphism, neither injective nor surjective. The kernel is $\mathrm{Ker}(\varphi) = 4\mathbb{Z} = \{4k \mid k \in \mathbb{Z}\}$. □

12.F.57. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

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.

12.4.1. 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.
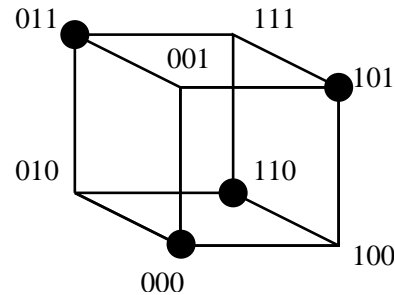
Moreover, the parity check code is unable to detect the error of interchanging a pair of adjacent bits, which often happens.

12.4.2. 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:

892

i) $\varphi : \mathbb{Q}^* \to \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{q}{p}$

ii) $\varphi : \mathbb{Q}^* \to \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{p^2}{q^2}$

iii) $\varphi : \mathbb{Q}^* \to \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{p^2+q^2}{pq}$

○

12.F.58. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathbb{C} \to \mathbb{R}, \varphi(a + bi) = a + b$,

ii) $\varphi : \mathbb{C} \to \mathbb{R}, \varphi(a + bi) = a$,

iii) $\varphi : \mathbb{C}^* \to \mathbb{R}^*, \varphi(a + bi) = a^2 + b^2$,

iv) $\varphi : \mathbb{C}^* \to \mathbb{R}^*, \varphi(c) = 2|c|$,

v) $\varphi : \mathbb{C}^* \to \mathbb{R}^*, \varphi(c) = |c|^3$,

vi) $\varphi : \mathbb{C}^* \to \mathbb{R}^*, \varphi(c) = 1/|c|$.

○

12.F.59. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathcal{GL}_2(\mathbb{R}) \to \mathbb{R}^*, \varphi(A) = |A|$

ii) $\varphi : \mathcal{GL}_2(\mathbb{R}) \to \mathbb{R}^*, \varphi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = a^2 + b^2$.

iii) $\varphi : \mathcal{GL}_2(\mathbb{R}) \to \mathbb{R}^*, \varphi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = ac + bd$.

○

12.F.60. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathbb{Z}_3 \to \mathbb{A}_4, \varphi([a]_3) = (1,2,4) \circ (1,3,2)^a \circ (1,4,2)$

ii) $\varphi : \mathbb{Z}_3 \to \mathbb{A}_4, \varphi([a]_3) = (1,2) \circ (1,3,2)^a$

○

12.F.61. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

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 $2r + 1$.

12.4.3. 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 $b_0 b_1 \ldots b_{k-1}$ is understood as the polynomial

$$m(x) = b_0 + b_1 x + \cdots + b_{k-1} x^{k-1}$$

over the field $\mathbb{Z}_2$. The encoded message should be another polynomial $v(x)$ of degree at most $n - 1$.

POLYNOMIAL CODES

Let $p(x) = a_0 + \cdots + a_{n-k} x^{n-k} \in \mathbb{Z}_2[x]$ 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 $m(x)$:

$$v(x) = r(x) + x^{n-k} m(x) =$$
$$= r(x) + q(x)p(x) + r(x) = q(x)p(x),$$

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.

i) $\varphi : \mathbb{Z} \to \mathbb{Z}$, $\varphi(a) = 2a$ iii) $\varphi : \mathbb{Z} \to \mathbb{Z}$,

ii) $\varphi : \mathbb{Z} \to \mathbb{Z}$, $\varphi(a) = 3|a|$

    $\varphi(a) = a + 1$ iv) $\varphi : \mathbb{Z} \to \mathbb{Z}$, $\varphi(a) = 1$

○

12.F.62. For each of the following formulas, decide whether it correctly defines a mapping φ. If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

i) $\varphi : \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \to \mathbb{Q}^*$, $\varphi((a,b,c)) = 2^a 3^b 12^c$

ii) $\varphi : \mathbb{Z}_3^* \times \mathbb{Z}_5 \to \mathbb{Z}_5$, $\varphi((a,b)) = b^a$

iii) $\varphi : \mathbb{Z}_2 \times \mathbb{Z} \to \mathbb{Z}$, $\varphi(([a]_2, b)) = b$

○

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 12.3.5). 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. □

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.

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

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

$$p(x) = 1 + x + \cdots + x^{n-1}$$

generates the $(n, 1)$-code of n-fold bit repetition. Dividing the polynomial $b_0 x^{n-1}$ by $p(x)$, gives the remainder $b_0(1 + \cdots + x^{n-2})$, so the corresponding codeword is $b_0 p(x)$.

12.4.4. 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

$$e(x) = x^i + x^j = x^i \left(1 + x^{j-i}\right)$$

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. □

12.4.5. 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

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 $|S_{\mathrm{id}}| = 84$, 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 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:

- $g = \mathrm{Id}$: $|S_g| = 3^9$.
- g is a rotation by $90°$ or $270°(= -90°)$: 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 $180°$: 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 $|S_g| = 3^6$.

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. □

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.[12]

12.4.6. 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:

---

[12]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.

By Burnside's lemma, the wanted number of colorings is equal to

$$\frac{1}{8}\left(3^9 + 2 \cdot 3^3 + 3^5 + 4 \cdot 3^6\right) = 2862. \qquad \square$$

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-semiaxis and which of the four adjacent vertices is on the positive y-semiaxis. Thus, the group has 24 elements. These are (besides the identity) rotations by $\pm 90°$ and $180°$ around axes going through opposite vertices, rotations by $180°$ around axes going through the centers of opposite edges, and finally rotations around $\pm 120°$ 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 $\pm 90°$ around an axis going through opposite vertices. Then, g fixes 3^2 colorings, and there are 6 such rotations.

- g is a rotation by $180°$ 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 $\pm 120°$. 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

$$\tfrac{1}{24}\left(3^8 + 6 \cdot 3^2 + 17 \cdot 3^4\right) = 333. \qquad \square$$

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.

Any injective linear mapping $g : (\mathbb{Z}_2)^k \to (\mathbb{Z}_2)^n$ 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. Evaluate it on the individual basis vectors $m_i(x) = x^{i-1}$, $i = 1, 2, 3$, to get

$$v_0 = (1 + x) + x^3,$$
$$v_1 = (x + x^2) + x^4,$$
$$v_2 = (1 + x + x^2) + x^5.$$

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 $\mathbb{I}_k$.

12.4.7. Theorem. *Let $g : (\mathbb{Z}_2)^k \to (\mathbb{Z}_2)^n$ be a linear code with generating matrix G, written in blocks as*

$$G = \begin{pmatrix} P \\ \mathbb{I}_k \end{pmatrix}.$$

Then, the mapping $h : (\mathbb{Z}_2)^n \to (\mathbb{Z}_2)^{n-k}$ with the matrix

$$H = \begin{pmatrix} \mathbb{I}_{n-k} & P \end{pmatrix}$$

has the following properties:

(1) $\operatorname{Ker} h = \operatorname{Im} g$,

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 $120°$ and $240°$, reflections, and of course the identity. By Burnside's lemma, the wanted number of necklaces is equal to

$$\frac{1}{36}\left(4084080 + 2 \cdot \binom{6}{3}\binom{3}{3} + 9 \cdot \binom{8}{4}\binom{4}{3}\right) = 113590.$$

$\square$

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.

(2) a received word u is a codeword if and only if $H \cdot u = 0$.

PROOF. The composition $h \circ g : (\mathbb{Z}_2)^k \to (\mathbb{Z}_2)^{n-k}$ is given by the product of matrices (computing over $\mathbb{Z}_2$):

$$H \cdot G = \begin{pmatrix} \mathbb{I}_{n-k} & P \end{pmatrix} \cdot \begin{pmatrix} P \\ \mathbb{I}_k \end{pmatrix} = P + P = 0.$$

Hence it is proved that $\operatorname{Im} g \subseteq \operatorname{Ker} h$. Since the first $n - k$ columns of H are basis vectors of $(\mathbb{Z}_2)^{n-k}$, the image $\operatorname{Im} h$ 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 12.3.10 can be used, thus obtaining

$$|\operatorname{Ker} h| \cdot |\operatorname{Im} h| = |(\mathbb{Z}_2)^n| = 2^n.$$

Therefore, the number of vectors in $\operatorname{Ker} h$ 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 $\operatorname{Im} g$ 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

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

12.4.8. 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 $h : (\mathbb{Z}_2)^n \to (\mathbb{Z}_2)^{n-k}$ corresponding to the parity check matrix has V as its kernel. Therefore, it induces the injective linear mapping $\tilde{h} : (\mathbb{Z}_2)^n/V \to (\mathbb{Z}_2)^{n-k}$. 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.

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

$$0, x^2 + x + 1, x^3 + x^2 + x, x^3 + 1, x^4 + x^3 + x^2, x^4 + x^3 + x + 1,$$
$$x^4 + x, x^4 + x^2 + 1$$

or

$$00000, 11100, 01110, 10010, 00111, 11011, 01001, 10101.$$

The basis vectors multiplied by $x^{5-3} = x^2$ yield mod (p):

$$x^2 \equiv x + 1,$$
$$x^3 = x.x^2 \equiv x(x+1) = x^2 + x \equiv 1,$$
$$x^4 \equiv x.$$

This means that the basis vectors are encoded as follows:

$$1 \mapsto x^2 + x + 1, \qquad\qquad 100 \mapsto 11100,$$
$$x \mapsto x^3 + 1, \qquad \text{i. e.} \qquad 010 \mapsto 10010,$$
$$x^2 \mapsto x^4 + x, \qquad\qquad 001 \mapsto 01001.$$

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

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}. \qquad \square$$

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 $a_0 a_1 \ldots a_6$, considered as the polynomial $a_0 + a_1 x + \cdots + a_6 x^6$, 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?

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
000000	**100000**	**010000**	**001000**
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
000100	**000010**	**000001**	**000110**
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.

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

$$
\begin{aligned}
1 + x + x^5 + x^6 \mapsto \quad &x^4 + x^5 + x^9 + x^{10} + x + 1 + x^2 + x + \\
&x^3 + x + x^2 + x + 1 = \\
= \quad &x^3 + x^4 + x^5 + x^9 + x^{10}.
\end{aligned}
$$

Thus, the code is 00011100011.

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. □

12.H.4. A 7-bit message $a_0 a_1 \ldots a_6$, considered as the polynomial $a_0 + a_1 x + \cdots + a_6 x^6$, 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 $\underbrace{0101}_{\text{redundancy}}\underbrace{1101011}_{\text{message}}$.

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.

5. 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.

12.5.1. 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 $f = \sum_\alpha a_\alpha x^\alpha \in \mathbb{K}[x_1, \ldots, x_n]$ can be viewed naturally as a mapping $f \colon \mathbb{K}^n \to \mathbb{K}$, defined by

$$
f(u_1, \ldots, u_n) := \sum_\alpha a_\alpha u^\alpha, \text{ where } u^\alpha = u_1^{\alpha_1} \cdots u_n^{\alpha_n}.
$$

In dimension $n = 1$, the equality $f(x) = 0$ describes only finitely many points of $\mathbb{K}$. In higher dimension, the similar equality $f(x_1, \ldots, x_n) = 0$ 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 $(x^2 + y^2)^3 - 4x^2 y^2 = 0$ 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* $x^2 - y^2 z = 0$, which, besides the part shown in the diagram, also includes the line $\{x = 0, y = 0\}$.

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. □

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 $x^4 \equiv x^3 + 1, x^5 \equiv x^3 + x + 1, \ldots, x^{12} \equiv x + 1$ and find out that the thirteenth it was flipped, and the original message was 01110111101.

Let us look at the exercise more thoroughly. Computing all powers of x, we obtain

$$x^4 \equiv x^3 + 1,$$
$$x^5 \equiv x^3 + x + 1,$$
$$x^6 \equiv x^3 + x^2 + x + 1,$$
$$x^7 \equiv x^2 + x + 1,$$
$$x^8 \equiv x^3 + x^2 + x,$$
$$x^9 \equiv x^2 + 1,$$
$$x^{10} \equiv x^3 + x,$$
$$x^{11} \equiv x^3 + x^2 + 1,$$
$$x^{12} \equiv x + 1,$$
$$x^{13} \equiv x^2 + x,$$
$$x^{14} \equiv x^3 + x^2,$$
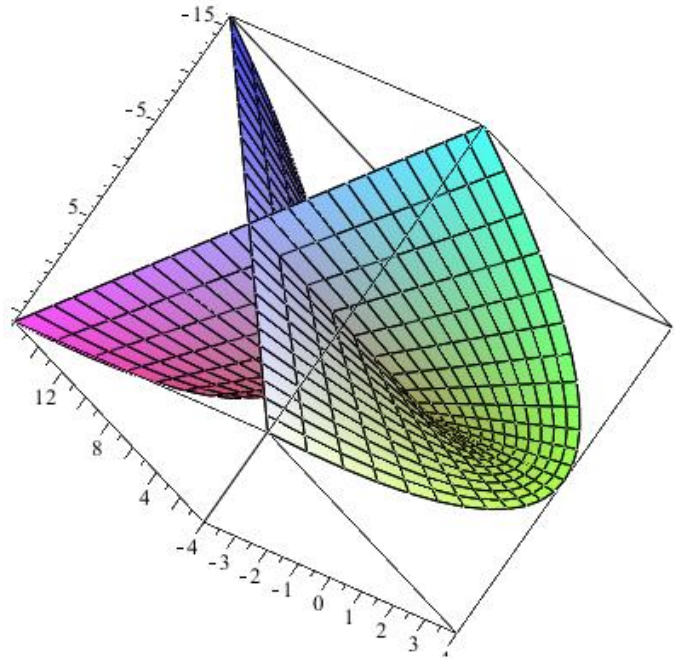
The diagram was drawn using the parametric description $x = uv$, $y = v$, $z = u^2$, whence the implicit description $x^2 - y^2 z = 0$ is easily guessed.

In the following illustration, there is the *Enneper surface* with parametrization $x = 3u+3uv^2-u^3$, $y = 3v+3u^2v-v^3$, $z = 3u^2 - 3v^2$.
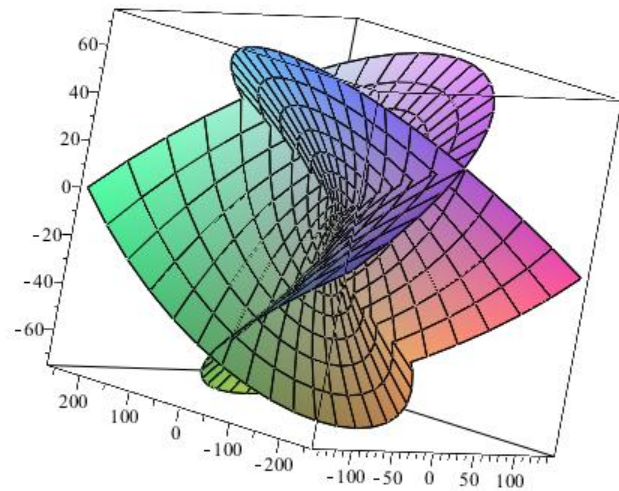
It is hard to imagine how to obtain the implicit description from this parametrization in hand. Nevertheless, there is an algorithm to eliminate the variables u and v from these three equations.

Quite a complex theory is needed to be developed for that. As usual, begin by formalization of the objects of interest.

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 $\qquad\square$

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

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

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 $\begin{pmatrix} \mathbb{I}_{n-k} & P \end{pmatrix}$, i. e., in our case

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

Let $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_n]$. The *affine variety* in $\mathbb{K}^n$ corresponding to the set of polynomials $f_1, \ldots, f_s$ is the set

$$\mathfrak{V}(f_1, \ldots, f_s) = \big\{ (a_1, \ldots, a_n) \in \mathbb{K}^n;$$
$$f_i(a_1, \ldots, a_n) = 0, i = 1, \ldots, s \big\}$$

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* $V = \mathfrak{V}(f_1, \ldots, f_s)$ *and* $W = \mathfrak{V}(g_1, \ldots, g_t) \subseteq \mathbb{K}^n$ *be affine varieties. Then,* $V \cup W$ *and* $V \cap W$ *are also affine varieties, where*

$$V \cap W = \mathfrak{V}(f_1, \ldots, f_s, g_1, \ldots, g_t),$$
$$V \cup W = \mathfrak{V}(f_i g_j) \quad \text{for } 1 \le i \le s, 1 \le j \le t.$$

In the following subsections, some questions which arise in the context of varieties are answered:

Q1. *Is the set* $\mathfrak{V}(f_1, \ldots, f_s)$ *empty?*
Q2. *Is the set* $\mathfrak{V}(f_1, \ldots, f_s)$ *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.

12.5.2. 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).

Multiplying the received word by the check matrix, we get the syndrome (error) of the word:

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} =$$

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

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 12.4.8). 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. □

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. ○

The variety $\mathfrak{V}(x + y + z - 1, x + 2y - z - 3)$ is a line (the intersection of two planes). If the system

$$\begin{array}{ccccccccc} x & + & y & + & z & - & 1 & = & 0, \\ x & + & 2y & - & z & - & 3 & = & 0, \end{array}$$

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 $\mathfrak{V}(f_1, \ldots, f_r) \subseteq \mathbb{K}^n$ is a set of rational functions $r_1, \ldots, r_n \in \mathbb{K}(t_1, \ldots, t_s)$ such that:
- the entire image of the mapping $r = (r_1, \ldots, r_n)$ is contained in $\mathfrak{V}(f_1, \ldots, f_r)$;
- $\mathfrak{V}(f_1, \ldots, f_r)$ is the minimal affine variety which contains these points $(x_1, \ldots, x_n)$.

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:

$$x = \frac{2t}{1 + t^2}, \ y = \frac{-1 + t^2}{1 + t^2},$$

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.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 $(0000, 0001, \ldots, 1111)$ by the generating matrix of the code. Thus, we get:

0110001,	1010010,	1100100,	0111000
1100011,	1010101,	0001001,	1011100
1101010,	0110110,	0001110,	1101101
1011011,	0000111,	0111111,	0000000.

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

1000001,	0110010,	0000100,	1011000,
0000011,	0110101,	1101001,	0111100,
0001010,	1010110,	1101110,	0001101,
0111011,	1100111,	1011111,	1100000

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. □

12.5.3. 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

$$f, g \in I \implies f + g \in I,$$
$$f \in I, h \in \mathbb{K} \implies f \cdot h \in I.$$

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 $I = \langle a_1, \ldots, a_n \rangle$. 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 $V = \mathfrak{V}(f_1, \ldots, f_s)$, set

$$\mathfrak{I}(V) := \big\{ f \in \mathbb{K}[x_1, \ldots, x_n];$$
$$f(a_1, \ldots, a_n) = 0, \forall (a_1, \ldots, a_n) \in V \big\}.$$

Lemma. *Let* $f_1, \ldots, f_s, g_1, \ldots, g_t \in \mathbb{K}[x_1, \ldots, x_n]$ *be polynomials. Then:*
(1) if $\langle f_1, \ldots, f_s \rangle = \langle g_1, \ldots, g_t \rangle$, *then*
$$V = \mathfrak{V}(f_1, \ldots, f_s) = \mathfrak{V}(g_1, \ldots, g_t);$$
(2) $\mathfrak{I}(V)$ *is an ideal, and* $\langle f_1, \ldots, f_s \rangle \subseteq \mathfrak{I}(V)$.

PROOF. If a point $a = (a_1, \ldots, a_n)$ lies in a variety $\mathfrak{V}(f_1, \ldots, f_s)$, then any polynomial of the form

$$f = h_1 f_1 + \cdots + h_s f_s$$

(i.e. any member of the ideal $I = \langle f_1, \ldots, f_s \rangle$) takes zero at a. In particular, this means that all the polynomials g_i take zero at a. Hence

$$\mathfrak{V}(f_1, \ldots, f_s) \subseteq \mathfrak{V}(g_1, \ldots, g_t).$$

The other inclusion can be proved similarly.

12.H.9. Consider the linear $(7, 4)$-code (i. e., the message has length 4) over $\mathbb{Z}_2$ defined by the matrix

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

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 □

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. □

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. □

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?

In order to verify the second proposition, choose $g, g' \in \mathfrak{I}(V), h \in \mathbb{K}[x_1, \ldots, x_n]$. 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 $\mathbb{K}[x_1, \ldots, x_n]$.

For any polynomial $f = h_1 f_1 + \cdots + h_s f_s \in \langle f_1, \ldots, f_s \rangle$ and a point $a \in V$, $f(a) = 0$, which proves the desired inclusion. □

The simplest examples are trivial varieties – a single point and the entire affine space:

$$\mathfrak{I}\big(\{(0, 0, \ldots, 0)\}\big) = \langle x_1, \ldots, x_n \rangle,$$
$$\mathfrak{I}(\mathbb{K}^n) = \{0\} \quad \text{for any infinite field } \mathbb{K}.$$

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 $I \in \mathbb{K}[x_1, \ldots, x_n]$ finitely generated?*

Q7. *Is there an algorithm which decides whether $f \in \langle f_1, \ldots, f_s \rangle$?*

Q8. *What is the precise relation between $\langle f_1, \ldots, f_s \rangle$ and $\mathfrak{I}\big(\mathfrak{V}(f_1, \ldots, f_s)\big)$?*

12.5.4. Dimension 1. Consider univariate polynomials first

$$f = a_0 x^n + a_1 x^{n-1} + \cdots + a_n, \quad \text{where } a_0 \neq 0.$$

The leading term of a polynomial is defined to be $LT(f) := a_0 x^n$. 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 $\langle f \rangle$.*

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. $\qquad\square$

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. $\qquad\square$

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.

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 $I = \langle f \rangle$. $\qquad\square$

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

$$\forall p \in \mathbb{K}[x] \colon p | r, g, \quad \text{so } p | f \text{ and } p | h.$$

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

$$GCD(f_1, \ldots, f_s) := GCD\big(f_1, GCD(f_2, \ldots, f_s)\big).$$

Lemma. *Let* $f_1, \ldots, f_s$ *be polynomials. Then,* $\langle GCD(f_1, \ldots, f_s) \rangle = \langle f_1, \ldots, f_s \rangle$.

PROOF. $GCD(f_1, \ldots, f_s)$ divides all the polynomials f_i. Hence the principal ideal $\langle GCD(f_1, \ldots, f_s) \rangle$ contains the ideal $\langle f_1, \ldots, f_s \rangle$. The other inclusion follows immediately from Bezout's identity. $\qquad\square$

Earlier, eight questions were formulated. Here are some answers for dimension 1:

- Since $\mathfrak{V}(f_1, \ldots, f_s) = \mathfrak{V}(GCD(f_1, \ldots, f_s))$, 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 $GCD(f_1, \ldots, f_s)$, except for the case when $GCD(f_1, \ldots, f_s) = 0$. This can happen only if $f_1 = f_2 = \cdots = f_s = 0$, and then the variety is the entire $\mathbb{K}$.

e., our parametrization of the circle) can be described as:

$$(t:1) \mapsto \left(\frac{2t}{1+t^2} : \frac{t^2-1}{t^2+1} : 1 \right) = \left(2t : t^2-1 : t^2+1 \right).$$

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 4.3.1): 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.

J. Elliptic curves

A *singular point* of a hypersurface in P^n, defined by a homogeneous polynomial

$$F(x_0, x_1, \ldots, x_n) = 0,$$

is such a point which satisfies $\frac{\partial F}{\partial x_i} = 0$ for $i = 1, \ldots, n$.

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 $(x^2+y^2)^3 - 4x^2y^2$ v $\mathbb{R}^2$:

- 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.
- $f \in \langle f_1, \ldots, f_s \rangle \iff GCD(f_1, \ldots, f_s) | f$.
- Denoting $\langle f \rangle := \mathfrak{I}(\mathfrak{V}(f_1, \ldots, f_s))$, then f and $GCD(f_1, \ldots, f_s)$ may differ only in root multiplicities.

12.5.5. 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 $f \in \mathbb{K}[x_1, \ldots, x_n]$ by polynomials $g_1, \ldots, g_s$ is to be an expression of the form

$$f = a_1 g_1 + \cdots + a_s g_s + r$$

where no term of the remainder r is divisible by the leading term of any of the polynomials g_i.

Try this with $f = x^2y + xy^2 + y^2$, $g_1 = xy - 1$, 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?

<div style="text-align:center">Monomial ordering</div>

A *monomial ordering* on $\mathbb{K}[x_1, \ldots, x_n]$ 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 $x_1 < x_2 < \ldots x_n$ 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 $x_1 > x_2 > \cdots > x_n$.

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* $<_{\text{grlex}}$ is defined by

$$\alpha >_{\text{grlex}} \beta \iff |\alpha| > |\beta| \quad \text{or}$$
$$|\alpha| = |\beta| \quad \text{and } \alpha >_{\text{lex}} \beta.$$
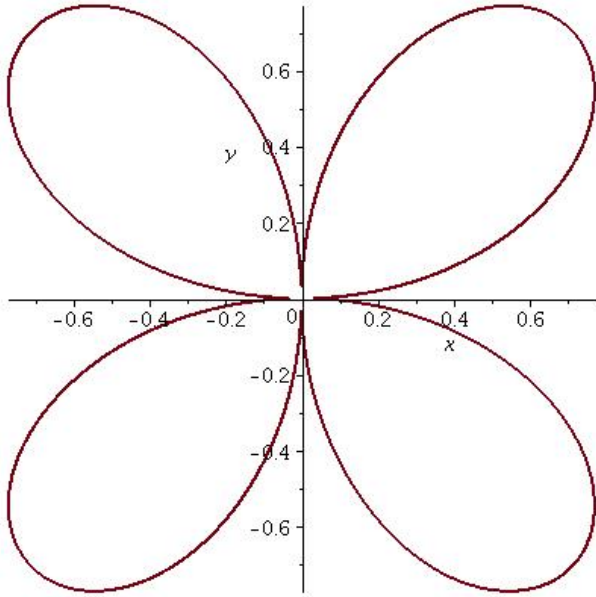
The cusp can be found on the curve in $\mathbb{R}^2$ given by $x^3 - y^2 = 0$.
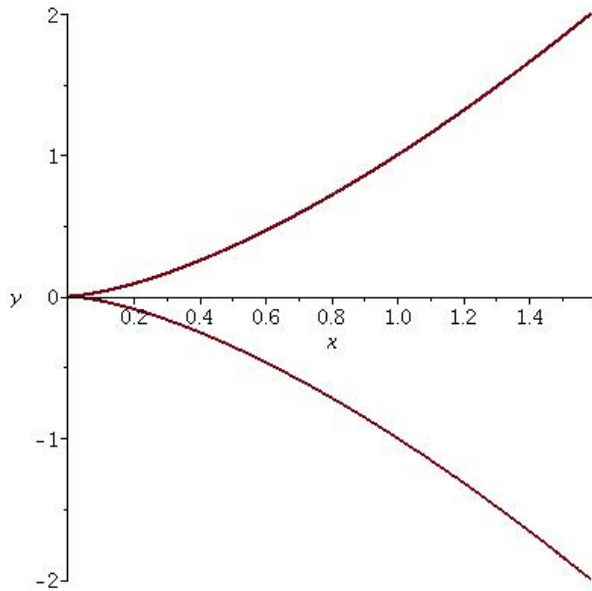
An *elliptic curve* $\mathcal{C}$ is the set of points in $\mathbb{K}^2$, where $\mathbb{K}$ is a given field, which satisfy an equation of the form

$$y^2 = x^3 + ax + b,$$

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 Δ 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.

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 $x^2yz^2 >_{\text{grlex}} xy^3z$, yet $x^2yz^2 <_{\text{grevlex}} xy^3z$.

Verify in detail that $>_{\text{lex}}, >_{\text{grlex}}, >_{\text{grevlex}}$ are monomial orderings!

12.5.6. Multivariate division with remainder. Consider a non-zero polynomial $f = \sum_{\alpha \in \mathbb{N}^n} a_\alpha x^\alpha$ in $\mathbb{K}[x_1, \ldots, x_n]$ 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 $f, g \in \mathbb{K}[x_1, \ldots, x_n]$ 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 $F = (f_1, \ldots, f_s)$ be an s-tuple of polynomials in $\mathbb{K}[x_1, \ldots, x_n]$. Then, every polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$ can be expressed as*

$$f = a_1 f_1 + \cdots + a_s f_s + r,$$

where $a_i, r \in \mathbb{K}[x_1, \ldots, x_n]$ for all $i = 1, 2, \ldots, s$. Moreover, either $r = 0$ or r is a linear combination of monomials none of which is divisible by any of $LT\, f_1, \ldots, LT\, f_s$, 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) $a_1 := 0, \ldots, a_s := 0, r := 0, p := f$
(2) `while` $p \neq 0$
  (a) $i := 1$
  (b) $d := \textit{false}$
  (c) `while` $i \leq s \wedge \text{not}\, d$

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 4.3.1) is $F(x, y, z) = 0$, where

$$(1) \qquad F(x, y, z) = y^2z - x^3 - axz^2 - bz^3.$$

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 γ 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 γ, all the three partial derivatives of F at the point $[\gamma, 0, 1]$ are zero. $\qquad \square$

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 4.3.1), and we define a point $O \in \mathcal{C}$ 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

(i) if $LT\, f_i | LT\, p$
   $a_i := a_i + LT\, p / LT\, f_i$
   $p := p - (LT\, p / LT\, f_i) \cdot f_i$
   $d := true$
   (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 $f = a_1 f_1 + \cdots + p + r$ 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. $\qquad \square$

In the ring $\mathbb{K}[x_1, \ldots, x_n]$, the following implication clearly holds:

$$f = a_1 f_1 + \cdots + a_s f_s + 0 \implies f \in \langle f_1, \ldots, f_s \rangle.$$

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 $I = \langle f_1, \ldots, f_s \rangle$ 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.

12.5.7. Monomial ideals. An ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ 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^α with $\alpha \in A$.

This means that all polynomials in I are of the form $\sum_{\alpha \in A} h_\alpha x^\alpha$, where $h_\alpha \in \mathbb{K}[x_1, \ldots, x_n]$.

Clearly, for a monomial ideal I $x^\beta \in I$ if and only if there exists an $\alpha \in A$ such that x^α divides x^β.

Lemma. *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ be a monomial ideal and $f \in \mathbb{K}[x_1, \ldots, x_n]$ 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 $h_\beta \in \mathbb{K}[x_1, \ldots, x_n]$. Each

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.
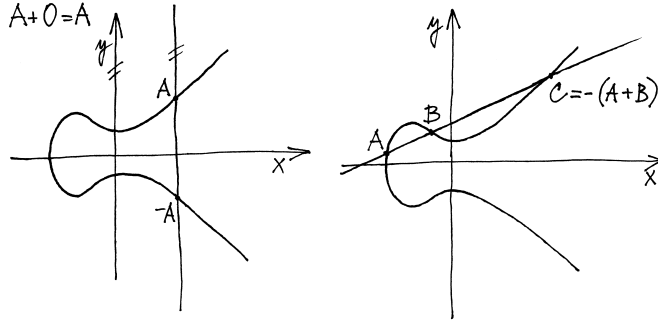
12.J.2. Prove that the above definition correctly defines an operation on the points of an elliptic curve.

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 points 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., $O + O = O$. □

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., $A = -B$.

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. □

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.

12.5.8. Theorem (Dickson's lemma). *Every monomial ideal $I = \langle x^\alpha,\ \alpha \in A \rangle \subseteq \mathbb{K}[x_1, \ldots, x_n]$ can be written in the form $I = \langle x^{\alpha_1}, \ldots, x^{\alpha_s} \rangle$, where $\alpha_1, \ldots, \alpha_s \in A$.*

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^β divides each monomial x^α 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 $x_1, \ldots, x_{n-1}, y$, and write monomials in the form $x^\alpha y^m$ where $\alpha \in \mathbb{N}^{n-1}$, $m \in \mathbb{N}$. Suppose that $I \subseteq \mathbb{K}[x_1, \ldots, x_{n-1}, y]$ is monomial, and define $J \subseteq \mathbb{K}[x_1, \ldots, x_{n-1}]$ 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, $J = \langle x^{\alpha_1}, \ldots, x^{\alpha_s} \rangle$. 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 $J_k \subseteq \mathbb{K}[x_1, \ldots, x_{n-1}]$ for $0 \le k \le m - 1$

$$J_k := \langle x^\beta;\ x^\beta y^k \in I_A \rangle.$$

Again, all the ideals J_k satisfy the induction hypothesis, so they can be expressed as

$$J_k = \langle x^{\alpha_{k,1}}, \ldots, x^{\alpha_{k,s_k}} \rangle.$$

It remains to show that I is generated by the following finite set of monomials:

$$x^{\alpha_1} y^m, \ldots, x^{\alpha_s} y^m,$$
$$x^{\alpha_{0,1}} y^0, \cdots, x^{\alpha_{0,s_0}} y^0,$$
$$\vdots$$
$$x^{\alpha_{m-1,1}} y^{m-1}, \ldots, x^{\alpha_{m-1,s_{m-1}}} y^{m-1}.$$

Consider a monomial $x^\alpha y^p \in I$. Either of the following cases occurs:

- $p \ge m$. Then, $x^\alpha \in J$, $k = p$, so one of $x^{\alpha_1} y^m, \ldots, x^{\alpha_s} y^m$ divides $x^\alpha y^p$.
- $p < m$. Then, analogously, $x^\alpha \in J_k$, and one of $x^{\alpha_{k,1}} y^k, \ldots, x^{\alpha_{k,s_k}} y^k$ divides $x^\alpha y^p$.

For a point $A \neq -B$, $A = [\alpha, \beta, 1]$ and a point $B \in \mathcal{C}$, $B = [\gamma, \delta, 1]$, we set

$$
\begin{aligned}
k &= \begin{cases} \frac{\beta - \delta}{\alpha - \gamma} & \text{for } A \neq B, \\ [5pt] \frac{3\alpha^2 + a}{2\beta} & \text{for } A = B, \end{cases} \\
\sigma &= k^2 - \alpha - \gamma, \\
\tau &= -\beta + k(\alpha - \sigma).
\end{aligned}
$$

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. $\qquad\square$

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 12.5.12, 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 12.5.6), 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 12.5.13, in order to get one, we must add the remainder polynomial $g_3 = y^3$ to g_1, g_2. Now, we calculate that

$$S(g_1, g_3) = y^3 g_1 - x^2 g_3 = 0$$

and

$$S(g_2, g_3) = y^2 g_2 - xg_3 = y^4 = yg_3.$$

Hence it follows by theorem 12.5.12 that g_1, g_2, g_3 is already a Gröbner basis. $\qquad\square$

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 from a Gröbner basis. We can obtain one by adding the polynomial $g_3 = x^3 - 2x^2$. Then, we have

$$S(g_1, g_3) = x^2 g_1 - yg_3 = 0$$

and

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. $\qquad\square$

12.5.9. 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 $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$, denote

$$LT\, I := \{ax^\alpha;\ \exists f \in I\colon\ LT\, f = ax^\alpha\}.$$

Clearly, $\langle LT\, I \rangle$ is a monomial ideal, so by Dickson's lemma, $\langle LT\, I \rangle = \langle LT\, g_1, \ldots, LT\, g_s \rangle$ for appropriate $g_1, \ldots, g_s \in I$.

Theorem. *Every ideal $I \in \mathbb{K}[x_1, \ldots, x_n]$ 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 $g_1, \ldots, g_s \in I$ such that $\langle LT\, I \rangle = \langle LT\, g_1, \ldots, LT\, g_s \rangle$.

Clearly, $\langle g_1, \ldots, g_s \rangle \subseteq I$. Choose any polynomial $f \in I$ and divide it by the s-tuple $g_1, \ldots, g_s$.

$$f = a_1 g_1 + \cdots + a_s g_s + r,$$

is obtained where no term of r is divisible by any of $LT\, g_1, \ldots, LT\, g_s$.

Since $r = f - a_1 g_1 - \cdots - a_s g_s$, $r \in I$, 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. $LT\, g_1, \ldots, LT\, g_s$. This contradicts the result of the multivariate division algorithm. Therefore, $r = 0$ and I is generated by $g_1, \ldots, g_s$. $\qquad\square$

12.5.10. 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 $g_1, \ldots, g_s$ of an ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ is called *Gröbner basis* if and only if $\langle LT\, I \rangle = \langle LT\, g_1, \ldots, LT\, g_s \rangle$.

Corollary. *Every ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ has a Gröbner basis. Every set of polynomials $g_1, \ldots, g_s \in I$ such that $\langle LT\, I \rangle = \langle LT\, g_1, \ldots, LT\, g_s \rangle$ 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.

$$S(g_2, g_3) = x^3 g_2 - y^2 g_3 = 2y^2 x^2 - x^5 =$$
$$= (4y + 2xy)g_1 - (x^2 + 2x + 4)g_3 + 8g_2. \qquad \square$$

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

$$g_1 = (\frac{1}{2}x + \frac{1}{4})f_1 + y^2 + z^2 - 1$$

and

$$y^2 + z^2 - 1 = (y - z + 1)f_2 + 2z^2 - 2z + \frac{1}{4}.$$

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 $I = \langle 2x - 1, y + z - 1, 8z^2 - 8z + 1 \rangle$. $\qquad \square$

12.K.4. Solve the following system of polynomial equations:

$$x^2 y - z^3 = 0,$$
$$2xy - 4z = 1,$$
$$z - y^2 = 0,$$
$$x^3 - 4yz = 0.$$

Solution. Using appropriate software, we can find out that the corresponding ideal

$$\langle x^2 y - z^3, 2xy - 4z - 1, z - y^2, x^3 - 4yz \rangle$$

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

$$x = 3u + 3uv^2 - u^3,$$
$$y = 3v + 3u^2 v - v^3,$$
$$z = 3u^2 - 3v^2.$$

This is the so-called Enneper surface, and it is depicted in the picture on page .

Denote the generators $f_i = \sum_j a_{ij} x_j + a_{i0}$. Consider the matrix $A = (a_{ij})$, where $i = 1, \ldots, s$ and $j = 0, \ldots, n$, 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 $g_1, \ldots, g_t$, where $t \leq s$.

Due to the performed steps, each f_i can be expressed as a linear combination $g_1, \ldots, g_t$, which means that

$$\langle f_1, \ldots, f_s \rangle = \langle g_1, \ldots, g_t \rangle.$$

Now, verify that these polynomials $g_1, \ldots, g_t$ form a Gröbner basis. Without loss of generality, assume that the variables are labeled so that $LM\, g_i = x_i$ for $i = 1, \ldots, t$. Any polynomial $f \in I$ can be written as

$$f = h_1 f_1 + \cdots + h_s f_s = h_1' g_1 + \cdots + h_t' g_t.$$

It is required that $LT\, f \in \langle LT\, g_1, \ldots, LT\, g_t \rangle$. That is, $LT\, f$ is divisible by one of $x_1, \ldots, x_t$. Suppose that f contains only the variables $x_{t+1}, \ldots, x_n$. However, then $h_1' = 0$, since x_1 is only in g_1 by the echelon form of B. Analogously, $h_2' = \cdots = h_t' = 0$, 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.

12.5.11. Theorem. *Let $G = \{g_1, \ldots, g_t\}$ be a Gröbner basis of an ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ and f a polynomial in $\mathbb{K}[x_1, \ldots, x_n]$. Then, there is a unique $r = \sum_\alpha a_\alpha x^\alpha \in \mathbb{K}[x_1, \ldots, x_n]$ such that:*

(1) no term of r is divisible by any of $LT\, g_1, \ldots, LT\, g_t$, i.e. $\forall \alpha \forall i \colon LT\, g_i \nmid a_\alpha x^\alpha$;

(2) $\exists g \in I \colon f = g + r$.

Proof. The algorithm for multivariate division produces

$$f = a_1 g_1 + \cdots + a_t g_t + r,$$

where r satisfies the condition (1). Select g as $a_1 g_1 + \cdots + a_t g_t$, 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 $LT\, g_1, \ldots, LT\, g_t$. There are two possibilities:

- $LM\, r \neq LM\, r'$. The one with the higher degree must be divisible by one of $LT\, g_1, \ldots, LT\, g_t$, 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 $LT\, g_1, \ldots, LT\, g_t$, which is again a contradiction.

It follows that $LT\, r = LT\, r'$ and the inductive argument shows that $r = r'$. $\qquad \square$

Solution. Applying the elimination procedure (e. g. int 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:

$$-59049z - 104976z^2 - 6561y^2 - 72900z^3 - 18954y^2z-$$
$$-23328z^4 + 32805z^2x^2 + 14580z^3x^2 + 3645z^4x^2 - 1296y^4z-$$
$$-16767y^2z^2 - 6156y^2z^3 - 783y^2z^4 + 39366zx^2 + 19683x^2-$$
$$-1296y^4 - 2430z^5 + 432z^6 + 108z^7 + 486z^5x^2 - 432y^4z^2+$$
$$+54y^2z^5 + 27z^6x^2 - 48y^4z^3 + 15y^2z^6 - 64y^6 - z^9.$$

□ 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, 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

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 $\overline{f}^G$ for the remainder in the division f/G, provided $G = (g_1, \ldots, g_s)$ is a Gröbner basis.

Corollary. *Let $G = \{g_1, \ldots, g_t\}$ be a Gröbner basis of an ideal $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ and f a polynomial in $\mathbb{K}[x_1, \ldots, x_n]$. Then, $f \in I$ if and only if the remainder $\overline{f}^G$ is zero.*

12.5.12. 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

$$\gamma := (\gamma_1, \ldots, \gamma_n), \quad \text{where } \gamma_i = \max\{\alpha_i, \beta_i\}.$$

The monomial x^γ 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 $\mathbb{R}[x, y]$ a with the $<_{\text{grlex}}$ ordering. Then, $\gamma = (4, 2)$ and

$$S(f, g) = \frac{x^4y^2}{x^3y^2}f - \frac{x^4y^2}{3x^4y}g = xf - \frac{1}{3}yg = -x^3y^3 + x^2 - \frac{1}{3}y^3,$$

which is a polynomial of degree 6.

Theorem. *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ be an ideal. Then, $G = \{g_1, \ldots, g_t\}$ 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 $c_1, \ldots, c_t \in \mathbb{K}$ and $\alpha_i + \text{multideg } g_i = \delta$ for a fixed δ whenever $c_i \neq 0$. If $\text{multideg } f < \delta$, then there are $c_{jk} \in \mathbb{K}$ such that*

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. □

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 α, β, γ. We want this robot to grasp, from above, an object which lies on the ground in distance x. What values should the angles α, β, γ 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 ground corresponds to the x-axis. It follows from elementary trigonometry that the total x-range of the robot at angels α, β, γ is equal to

$$x = \sin\alpha + \sin(\alpha + \beta) + \sin(\alpha + \beta + \gamma).$$

Similarly, the range of the robot in the vertical direction is

$$y = \cos\alpha + \cos(\alpha + \beta) + \cos(\alpha + \beta + \gamma).$$

$$\sum_{i=1}^{t} c_i x^{\alpha_i} g_i = \sum_{j,k=1}^{t} c_{jk} x^{\delta - \gamma_{jk}} S(g_j, g_k),$$

where $x^{\gamma_{jk}} = LCM(LM\,g_j, LM\,g_k)$, and the degree of each monomial $x^{\delta - \gamma_{jk}} S(g_j, g_k)$ is less than δ.

PROOF. Let $d_i := LC\,g_i$ and $p_i = x^{\alpha_i} g_i/d_i$. Clearly, $c_i d_i = LC(c_i x^{\alpha_i} g_i)$ and $LC\,p_i = 1$. Since multideg$(c_i x^{\alpha_i} g_i) = \delta$ and multideg $f < \delta$, it follows that $\sum_{i=1}^{t} c_i d_i = 0$. Express f as a combination of S-polynomials:

$$f = \sum_{i=1}^{t} c_i d_i p_i = c_1 d_1(p_1 - p_2) + (c_1 d_1 + c_2 d_2)(p_2 - p_3) +$$
$$+ \cdots + (c_1 d_1 + \cdots + c_{t-1} d_{t-1})(p_{t-1} - p_t) +$$
$$+ \underbrace{(c_1 d_1 + \cdots + c_t d_t)}_{0} p_t.$$

Each difference $p_j - p_k$ can be expressed in terms of S-polynomials

$$\frac{x^\delta}{d_j x^{\delta - \alpha_j}} g_j - \frac{x^\delta}{d_k x^{\delta - \alpha_k}} g_k = x^{\delta - \gamma_{jk}} \left(\frac{x^{\gamma_{jk}}}{LT\,g_j} g_j - \frac{x^{\gamma_{jk}}}{LT\,g_k} g_k \right) =$$
$$= x^{\delta - \gamma_{jk}} S(g_j, g_k)$$

Now the individual coefficients c_{jk} can be derived easily from these equalities. □

Now follows the proof of the theorem. The "$\Longrightarrow$" implication follows directly from the corollary of subsection 12.5.11.

For the reverse implication, consider a non-zero polynomial $f \in I$. It must be shown that under the implication assumption, $LT\,f \in \langle LT\,g_1, \dots LT\,g_t \rangle$. 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, multideg $f \leq \delta$. Let polynomials $h_1, \dots, h_t$ be chosen so that δ is as small as possible. Since this is a monomial ordering, which, in particular, is a well-ordering, such a δ exists.

It is necessary to prove that multideg $f = \delta$. Write

$$f = \sum_{m_i = \delta} h_i g_i + \sum_{m_i < \delta} h_i g_i =$$
$$= \sum_{m_i = \delta} (LT\,h_i) g_i + \sum_{m_i = \delta} (h_i - LT\,h_i) g_i + \sum_{m_i < \delta} h_i g_i.$$

Clearly, the degrees of all the terms of the second and third sums are less than δ. Admitting that multideg $f < \delta$, it follows that

$$\text{multideg}\left(\sum_{m_i = \delta} (LT\,h_i) g_i \right) < \delta.$$

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

$$\sin\alpha + \sin(\alpha+\beta) = x,$$
$$\cos\alpha + \cos(\alpha+\beta) - 1 = 0.$$

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:

$$s_1 + s_1 c_2 + c_1 s_2 - x = 0,$$
$$c_1 + c_1 c_2 - s_1 s_2 - 1 = 0,$$
$$s_1^2 + c_1^2 - 1 = 0,$$
$$s_2^2 + c_2^2 - 1 = 0.$$

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

$$(2c_2 + 1 - x^2, 2c_1(1+x^2) - 2s_2 x - 1 - x^2, 2s_1(1+x^2)+$$

$$+2s_2 - x - x^3, 4s_2^2 - 3 - 2x^2 + x^4),$$

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.

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, $\operatorname{multideg}(p_{ijk}g_i) \le \operatorname{multideg} S(g_j, g_k)$. 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,

$$\operatorname{multideg}(q_{ijk}g_i) \le \operatorname{multideg}\big(x^{\delta-\gamma_{jk}} S(g_j, g_k)\big) < \delta.$$

Substitution yields

$$\sum_{m_i=\delta} (LT\, h_i)g_i = \sum_{j,k} c_{jk}\left(\sum_{i=1}^{t} q_{ijk} g_i\right) =$$

$$= \sum_{i=1}^{t}\left(\sum_{j,k} c_{jk} q_{ijk}\right) g_i.$$

At the same time,

$$\operatorname{multideg}\left(\sum_{j,k} c_{jk} q_{ijk} g_i\right) < \delta \quad \text{for } i = 1, \dots, t.$$

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 δ. This contradicts the minimality of δ, and so $\operatorname{multideg} f = \delta$, whence $LT\, f \in \langle LT\, g_1, \dots, LT\, g_t \rangle$. So G is a Gröbner basis. $\qquad\square$

12.5.13. 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 $G \ne G'$
    (a) $G' := G$
    (b) $\forall p, q \in G' : p \ne q$ do
        (i) $s := \overline{S(p,q)}^{\,G'}$
        (ii) if $s \ne 0$
            $G := G \cup \{s\}$

$x = 1$ $x = \sqrt{3}$

a)   b)

$x = 1{,}5$ $x = 1{,}5$

□

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.,

$$f_1(x, y, z) = \left(2x, \frac{1}{2}y, z\right) \text{ and } f_2(x, y, z) = \left(2x, \frac{y-1}{2}, x+z\right).$$

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 nontrivial 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 $I_1 = LT(F) \subseteq I_2 \subseteq \cdots \subseteq I_k \subseteq \cdots$. 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, $I_k = I_{k+1} = \dots$.[13]

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.

12.5.14. 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 $\langle LT(G \setminus \{p\})\rangle = \langle LT\,G\rangle$. Hence the proposition follows immediately. □

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 $<_{\text{grlex}}$, $I = \langle f_1, f_2\rangle = \langle x^3 - 2xy, x^2y - 2y^2 + x\rangle$. 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:

---

[13]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".

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

$$F_1 : \quad x \mapsto ux \quad y \mapsto vy \quad z \mapsto z,$$

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 $F_1(xy - XY) = xy - XY$ ans $F_1(z - Z) = z - Z$, 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

$$I_2 = \langle uxv(y + 1) - ux - XY, (u - 1)x + z - Z, uv - 1 \rangle$$

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 $z = ab$. $\qquad \square$

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 $G = \{g_1, \ldots, g_s\}$, $G' = \{g_1', \ldots, g_t'\}$ 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, $m = a_1\, LT\, g_1' + \cdots + a_t\, LT\, g_t'$ for appropriate polynomials $a_1, \ldots, a_t$. 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 $LT\, g_1, \ldots, LT\, g_s$, 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. $\qquad \square$

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.

12.5.15. Theorem. *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ 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

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

$$x^3 - 2xy = 0,$$
$$x^2y + x - 2y^2 = 0.$$

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

$$yf_1 - xf_2 = -x^2.$$

The resulting basis can be reduced using division of the polynomials f_1, f_2 by x^2. This leads to the basis

$$\left(xy, x - 2y^2, x^2\right).$$

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

$$\left(x - 2y^2, y^3\right),$$

and this is a Gröbner basis: by the naive algorithm (see 12.5.13), it suffices to verify that the polynomial

$$S\left(x - 2y^2, y^3\right) = y^3\left(x - 2y^2\right) - xy^3 = -2y^5$$

gives zero remainder with respect to the basis $\left(x - 2y^2, y^3\right)$, even for any polynomial ordering. Clearly, the solution of the system is the point $(0, 0)$. □

12.K.10. Consider the following system of polynomial equations:

$$x^2yz^2 + x^2y^2 + yz - xyz^2 - z^2 = 0,$$
$$x^2y + z = 0,$$
$$xyz + z + 1 = 0.$$

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.

$$x^2y^2 + x^2yz^2 - xyz^2 + yz - z^2 = \left(y + z^2\right)\left(x^2y + z\right) -$$
$$- y(xyz + z + 1) - z^3 + z.$$

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. □

12.5.16. 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 12.5.13, 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

Hence $z = 0, \pm 1$. Then, e. g.,

$$0 = z\left(x^2 y + z\right) - x(xyz + z + 1) =$$
$$= z^2 - zx - x.$$

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 $\left(\frac{1}{2}, -4, 1\right)$. $\square$

12.K.11. Using Gröbner basis, solve the polynomial system

$$x^2 + y + z = 1,$$
$$x + y^2 + z = 1,$$
$$x + y + z^2 = 1.$$

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.

$$f_3 = 2yz^2 + z^4 - z^2.$$

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

$$g_4 = z^6 - 4z^4 + 4z^3 - z^2 =$$
$$= z^2(z-1)^2\left(z^2 + 2z - 1\right).$$

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 $(1, 0, 0), (0, 1, 0), (0, 0, 1), \left(-1 + \sqrt{2}, -1 + \sqrt{2}, -1 + \sqrt{2}\right)$, $\left(-1 - \sqrt{2}, -1 - \sqrt{2}, -1 - \sqrt{2}\right)$. $\square$

12.K.12. Solve the following system of polynomial equations in $\mathbb{R}$

$$x^2 - 2xz - 4,$$
$$x^2 y^2 z + yz^3,$$
$$2xy^2 - 3z^3.$$

Solution. The basis suitable for variable elimination is the Gröbner basis for the lexicographic monomial ordering with

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 $g_1, \ldots, g_k$ of the basis, that is, subsets S in the space $(\mathbb{K}[x_1, \ldots, x_n])^k$. 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 $\mathfrak{V}(g_1, \ldots, g_k)$.

12.5.17. Elimination of variables. We finish this chapter by an application of the above algorithms.

Consider the ring $\mathbb{K}[x_{p+1}, \ldots, x_n]$ to be a subring of $\mathbb{K}[x_1, \ldots, x_n]$. These are polynomials with no occurrence of the variables $x_1, \ldots, x_p$. It is a subring, but not an ideal.

ELIMINATION IDEALS

Let $I = \langle f_1, \ldots, f_s \rangle \subseteq \mathbb{K}[x_1, \ldots, x_n]$. For $p = 1, \ldots, n$, define

$$I_p := I \cap \mathbb{K}[x_{p+1}, \ldots, x_n].$$

This set is called the *p-th elimination ideal*. Note that I_p is an ideal only in $\mathbb{K}[x_{p+1}, \ldots, x_n]$.

On the level of polynomial equations, I_p contains all equations which are consequences of the system $f_1 = 0, \ldots, f_s = 0$ and which contain only the variables $x_{p+1}, \ldots, x_n$.

Theorem (Elimination theorem). *Let $I \subseteq \mathbb{K}[x_1, \ldots, x_n]$ be an ideal and $G = \{g_1, \ldots, g_m\}$ a Gröbner basis of I with respect to $<_{lex}$. Let the variables be ordered as $x_1 >_{lex} x_2 >_{lex} \cdots$. Then, for each $p = 0, \ldots, n$, $G_p := G \cap \mathbb{K}[x_{p+1}, \ldots, x_n]$ 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 $G_p = \{g_1, \ldots, g_r\}$. 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

$$f = h_1 g_1 + \cdots + h_r g_r.$$

To do this, perform multivariate division by the original Gröbner basis G. Since $f \in I$, it follows that $\overline{f}^G = 0$, i.e.

$$f = h_1 g_1 + \cdots + h_r g_r + h_{r+1} g_{r+1} \cdots + h_m g_m.$$

Each of the polynomials $g_{r+1}, \ldots, g_m$ must contain at least one of the variables $x_1, \ldots, x_p$, otherwise it would lie in G_p. By the properties of the lexicographic ordering, this variable must also be contained in $LT\, g_{r+1}, \ldots, LT\, g_m$.

Recall the individual steps of the algorithm for multivariate division and the fact that f contains no monomial with any of $x_1, \ldots, x_p$. Then $h_{r+1} = \cdots = h_m = 0$. thus verifing that $f \in \langle G_p \rangle$.

$x > y > z$. Using Maple, we can find the basis

$$144z^5 + 35z^7 + 12z^9,$$
$$23z^6 + 12z^8 + 44yz^4,$$
$$yz^3 + 3z^5 + 4zy^2, 9z^4 + 4y^3,$$
$$-8y^2 - 6z^4 + 3xz^3, 2xy^2 - 3z^3,$$
$$x^2 - 2xz - 4.$$

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 $y = 0, x = \pm 2$. □

12.K.13. Solve the following system of polynomial equations in $\mathbb{R}$:

$$xy + yz - 1,$$
$$yz + zw - 1,$$
$$zw + wx - 1,$$
$$wx + xy - 1.$$

Solution. In this case, it is a good idea to take the graded lexicographic ordering with $w > x > y > z$. Using the algorithm 12.5.13 or appropriate software, we find the corresponding Gröbner basis

$$(x - z, w - y, 2yz - 1).$$

Thus, the system is satisfied by exactly the points $\left(\frac{1}{2t}, t, \frac{1}{2t}, t\right)$ for an arbitrary $t \in \mathbb{R}$ except zero. □

12.K.14. Solve the following system of polynomial equations in $\mathbb{R}$:

$$x^2 + yz + x,$$
$$z^2 + xy + z,$$
$$y^2 + xz + y.$$

Solution. Using the algorithm 12.5.13 or appropriate software, we find the corresponding Gröbner basis for the lexicographic monomial ordering with $x > y > z$, consisting of six polynomials:

$$z^2 + 3z^3 + 2z^4,$$
$$z^2 + z^3 + 2yz^2 + 2yz^3,$$
$$y - yz - z - z^2 - 2yz^2 + y^2,$$
$$yz + z + z^2 + 2yz^2 + xz,$$
$$z^2 + xy + z, x^2 + yz + x.$$

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

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 $1 \le i < j \le r$, 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. □

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.

12.5.18. 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

$$x_1 = f_1(u_1, \ldots, u_k), \ldots, x_n = f_n(u_1, \ldots, u_k),$$

the reduced Gröbner basis of the ideal

$$\langle x_1 - f_1, \ldots, x_n - f_n \rangle$$

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.

$$x_i = \frac{f_i(t_1, \ldots, t_m)}{g_i(t_1, \ldots, t_m)},$$

it is perhaps natural to think of substituting the ideal

$$\langle x_1 g_1 - f_1, \ldots, x_n g_n - f_n \rangle$$

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 $I_2 = \langle z(x^2 y - z^3) \rangle$. However, the correct result is $\mathfrak{V}(x^2 y - z^3)$. 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

$(0,0,0), (-1,0,0), (0,-1,0), (0,0,-1)$ and $(-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$.

□

is included in $W = \mathfrak{V}(g_1, \ldots, g_n)$. Instead, perceive the parametrization F as a mapping $F : (\mathbb{K}^k - W) \to \mathbb{K}^n$. For the implicit situation, use the ideal

$$I = \langle g_1 x_1 - f_j, \ldots, g_n x_n - f_n, 1 - g_1 \cdots g_n y \rangle \subseteq$$
$$\subseteq \mathbb{K}[y, t_1, \ldots, t_m, x_1, \ldots, x_n],$$

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

12.A.2. $A' \wedge C'$.

12.A.3. $(A \text{ NAND}(B \text{ NAND } B))$.

12.A.31. E. g., $\{1, 2, 3, 12, 18\}$.

12.A.33. There are six isomorphism classes. In three of them ("Y, dual Y, and pentagon"), there is an element incomparable with two other ones, yielding 5! partial orders. In the other three ("X, house, and dual house"), there are two pairs of different incomparable elements, thus yielding only 5!/4 partial orders. Altogether, there are 450 partial orders.

12.B.7. 31.

12.B.8. 45.

12.B.9. 63.

12.B.10. 33.

12.C.1. Suppose the contrary, i. e., that the polynomial is a product of two polynomials with integer coefficients. Use induction to prove that all the coefficients of one of these polynomials are divisible by p (begin with the absolute term). However, then the leading coefficient of $f(x)$ is also divisible by p.

12.C.12. Over $\mathbb{R}$: $(x - 1)(2x^2 - x + 1)^2$, over $\mathbb{C}$: $(x - 1)\left(x - \frac{1 \pm i\sqrt{7}}{4}\right)^2$.

12.C.13. Over $\mathbb{R}$: $(x + 1)(x^2 + x + 2)^2$, over $\mathbb{C}$: $(x + 1)\left(x + \frac{1 \pm i\sqrt{7}}{4}\right)^2$.

12.C.14. Over $\mathbb{R}$: $(x^2 - 3x + 2)^2$, over $\mathbb{C}$: $\left(x - 1 + \sqrt{2}i\right)^2\left(x - 1 - \sqrt{2}i\right)^2$.

12.C.15. $x^5 + x^2 + 2x + 1 = (x^2 + 1)(x^3 + 2x + 1)$.

12.C.16. $x^4 + 2x^3 + 2$ is irreducible. It has no roots and it cannot be written as a product of two quadratic polynomials (this must be verified!).

12.E.3. i) not even a groupoid (the operation is not closed on the given set), ii) a non-commutative monoid, iii) a commutative group, iv) a non-commutative group, v) a commutative group, vi) a commutative monoid.

12.E.4. Since multiplication of complex numbers is commutative, it must remain such on any subset as well. The particular cases are:

 i) a monoid,
 ii) a group,
 iii) a group.

12.E.5.

 i) a commutative semigroup, but not a monoid;
 ii) a non-commutative groupoid, but not a semigroup;
 iii) a non-commutative groupoid, but not a semigroup;
 iv) a non-commutative semigroup, but not a monoid;
 v) a commutative monoid, but not a group;
 vi) a commutative monoid, but not a group;
 vii) a non-commutative group.

12.E.8. It is a non-commutative group.

12.F.1.

 i) $(1, 3, 5, 7, 2, 4, 6)$
 ii) $(1, 3, 2) \circ (4, 6, 5), (1, 4, 2, 5, 3, 6), (1, 5, 2, 6, 3, 4), (1, 6, 2, 4, 3, 5)$
 iii) none exists

12.F.4. Due to the parity, no such permutation exists.

12.F.14.

i) Yes	v) Yes
ii) No	vi) No
iii) Yes	vii) No
iv) No	viii) No

12.F.15.

 i) Yes

ii) No

12.F.16. $m = 10$. (Note that for $m = 8$ and $m = 12$, the resulting groups have the desired number of elements but are not isomorphic to $\mathbb{Z}_5^\times$.)

12.F.42. This is generally not true. Consider e. g. $\mathbb{S}_n/\mathbb{A}_n \sim Z_2$, $n \geq 3$.

12.F.47. The subgroup has four elements; the remaining one is the reflection with respect to the plane that is perpendicular to the former one and contains the axis of the rotation (it is isomorphic to the Klein group $\mathbb{Z}_2 \times \mathbb{Z}_2$). It is not normal.

12.F.57.
  i) An isomorphism.
  ii) A homomorphism, neither surjective, nor injective.
  iii) Not a homomorphism.

12.F.58.
  i) A surjective homomorphism, not injective.
  ii) A surjective homomorphism, not injective.
  iii) A homomorphism, neither surjective, nor injective.
  iv) Not a homomorphism.
  v) A homomorphism, neither surjective, nor injective.
  vi) A homomorphism, neither surjective, nor injective.

12.F.59.
  i) A surjective homomorphism, not injective.
  ii) Not a homomorphism.
  iii) Not a homomorphism.

12.F.60.
  i) An injective homomorphism, not surjective.
  ii) Not a correct definition, since the result does not lie in the specified codomain $\mathbb{A}_4$.

12.F.61.
  i) An injective homomorphism, not surjective.
  ii) Not a homomorphism.
  iii) Not a homomorphism.
  iv) Not a homomorphism.

12.F.62.
  i) A homomorphism, neither injective, nor surjective.
  ii) Not a mapping.
  iii) A surjective homomorphism, not injective.

12.G.5. $\frac{1}{36} \left(\frac{18!}{(6!)^3} + 2 \cdot 3! + 2 \cdot \frac{6!}{(2!)^3} + \frac{9!}{(3!)^3} + 18 \frac{9!}{(3!)^3} \right) = 477368.$

12.G.6. $\frac{1}{48} \left(\frac{24!}{(8!)^3} + \frac{12!}{(4!)^3} + 2 \frac{6!}{2^3} + 4 \cdot 3! + 24 \frac{12!}{(4!)^3} \right) = 197216213.$

12.G.7. 7.

12.H.2.
$$
G = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad
H = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}.
$$

12.H.7. 110.

Combinatorial methods, graphs, and algorithms

Do we often prefer thinking in pictures?
– yes, but we can compute discrete things only. . .
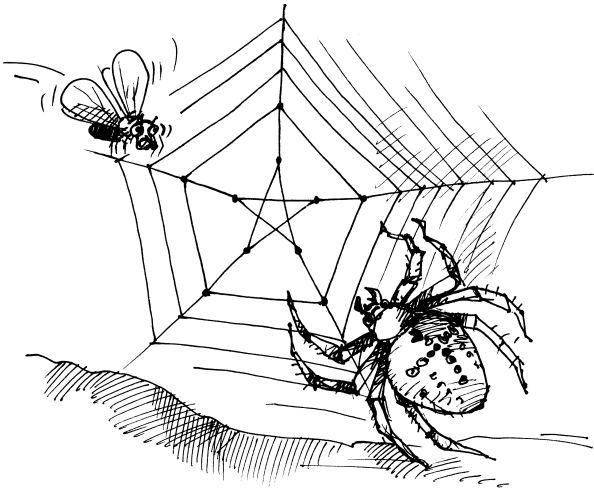
A. Fundamental concepts

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.

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?
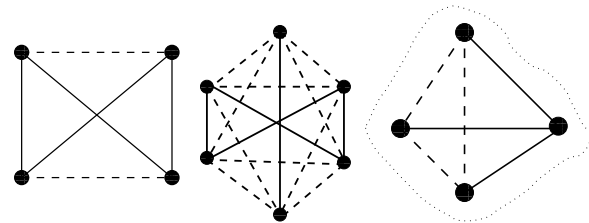
There is no such triangle in the left-hand diagram with four vertices. The example of a regular pentagon, in which all its outside edges are full, while all its diagonals are dashed (draw a picture!), shows that at least six vertices are required.

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.

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 with all degrees and $V_2 \subset V$- set of vertices with even degrees. Then,

$$2|E| = \sum_{v \in V} d(v) = \sum_{v \in V_1} d(v) + \sum_{v \in V_2} d(v), and$$

$$\sum_{v \in V_1} d(v) = 2|E| - \sum_{v \in V_2} d(v),$$

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 $v_1 \ldots v_k$ 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 $\{v_2, \ldots, v_k\}$ 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 $v_1 \ldots v_i v_1$ 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. □

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:
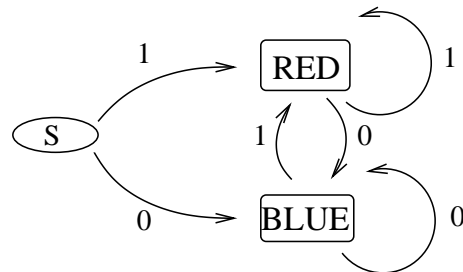
The third vertex which has only two outgoing arrows represents the beginning of the system (before the first bit is sent).

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.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. □

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 $v_{i_1} \ldots v_{i_k}$ 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. □

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. □

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

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 $K_n = \left(V, \binom{V}{2} \right)$ 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

odd degree, we are be 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. □

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 of minimal length, if there is a shorter path it will be shorter than this one, so the triangle inequality will be satisfied. □

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. □

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

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 $(v_0, \ldots, v_n)$ so that $E = \{e_1, \ldots, e_n\}$, where $e_i = \{v_{i-1}, v_i\}$ for all $i = 1, \ldots, n$. 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.
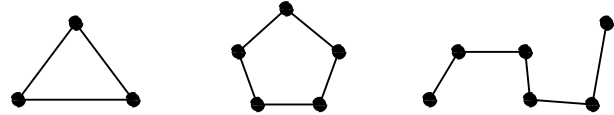
Another type of graph is the *complete bipartite graph*. Its vertices can be coloured with two (distinct) colours. All possible edges between vertices of different colours are present, but no other edges. Such a graph is denoted by $K_{m,n}$, where m and n are the numbers of vertices of particular colours. The diagram below illustrates the graphs $K_{1,3}$, $K_{2,3}$, and $K_{3,3}$.
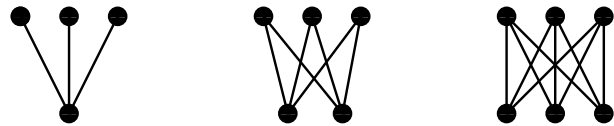
Another interesting example of a graph is the *hypercube H_n* in dimension n, whose vertices are the integers $0, \ldots, 2^n - 1$ and whose edges join those pairs of vertices whose binary expansions differ by exactly one bit. The following diagram depicts the hypercube H_4, with labels of the vertices 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).
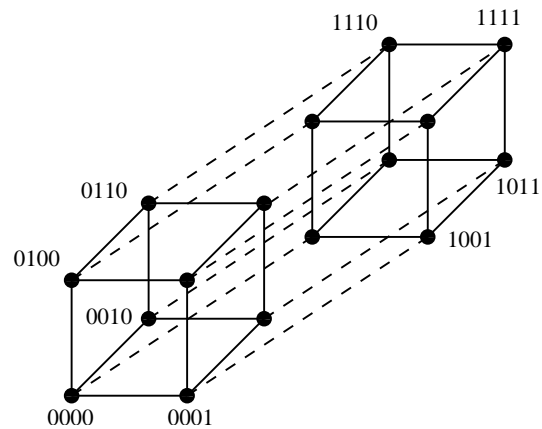
Here are two more examples. The first is the *cycle ladder graph CL_n* with $2n$ vertices. This consists of two cycle graphs C_n whose vertices are connected by edges according to their order in the cycles. The second is the *Petersen graph*.

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 $v_{i_1} \ldots v_{i_k} v_{i_1}$ in Q_k. In the first copy of Q_k that we take a cycle $0v_{i_1} \ldots 0v_{i_k} 0v_{i_1}$, and in the second copy $1v_{i_1} \ldots 1v_{i_k}, 1v_{i_1}$. Consider path $0v_{i_1} \ldots 0v_{i_k}, 1v_{i_k} 1v_{i_{k-1}} 1v_{i_1} 0v_{i_1}$ which forms a Hamiltonian cycle on Q_{k+1}. □

13.A.13. Show that a tree with n vertices has exactly $n1$ edges

Solution. We proceed by induction. Suppose that every tree with k vertices has precisely $k1$ 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 $k1$ edges. But to convert T_0 to T we need to add one edge and one vertex, so T also satises the formula for the number of edges. □

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 rst time, and this must occur eventually, since we know

This is somewhat similar to CL_5, yet it is actually the simplest counterexample for many propositions about graphs.
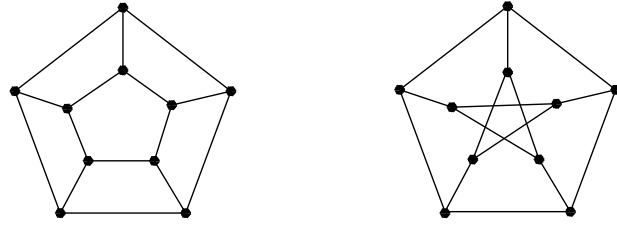
13.1.4. Morphisms of graphs. Mappings between the sets of vertices or edges which respect the considered structure are of great importance in graph theory. It is enough to consider mappings between the vertices only.

<center>MORPHISMS OF GRAPHS</center>

Definition. Let $G = (V, E)$ and $G' = (V', E')$ be two given graphs. A *morphism* (or *homomorphism*) $f : G \to G'$ is a mapping $f_V : V \to V'$ 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 $f : G \to G'$ 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:

<center>927</center>

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. $\qquad \square$

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 $P = v_{i_1} \ldots v_i$ and $Q = v_{j_1} \ldots v_{j_k}$ 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 $R = u_{i_1} \ldots u_{i_s}$, where $u_{i_1} = v_{i_1}$ and $u_{i_s} = v_{j_k}$. Path R coincides with P on some vertices, then it deviates from P. May be, returns to P again, then deviates from P for the last time and connects to Q. Denote this fragment of R by $v_{i_r} = u_{i_m} u_{i_{m+1}} \ldots u_{i_{m+t}} = v_{j_\ell}$, where $t \geq 1$. Let P' be path $v_{i_1} \ldots v_{i_r}$, $T' = u_{i_m} \ldots u_{i_{m+t}}$ and Q' be the longest part of Q going forward or backwards from v_{j_ℓ}, which length is obviously, at least, $k + 1$. $\qquad \square$

13.A.16. Solution.

$\qquad \square$

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 (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

A *walk of length n* in a graph G is a morphism $s : P_n \to G$. 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 $p : P_n \to G$ such that p is an injective mapping. The images of the vertices $v_0, \ldots, v_n$ from P_n are pairwise distinct.

A *cycle of length n* in a graph G is any morphism $c : C_n \to G$ 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 $(v_0, e_1, v_1, \ldots, e_n, v_n)$, where $e_i = \{v_{i-1}, v_i\}$ for $i = 1, \ldots, n$.

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.

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 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.

of the graph, so they will never meet at a party (in particular, not at the upcoming one). $\qquad\square$

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. $\qquad\square$

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 $1 + \cdots + 9 = \frac{1}{2} \cdot 9 \cdot 10 = 45$, 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.

$$(1, 1, 1, \underline{2}, \underline{2}, \underline{3}, \underline{4}, \underline{5}, \mathbf{5}) \longleftrightarrow (1, 1, 1, \underline{1}, \underline{1}, \underline{2}, \underline{3}, \mathbf{4}) \longleftrightarrow$$

$$\longleftrightarrow (1, 1, 1, 0, 0, 1, 2) \longleftrightarrow (0, 0, 1, 1, \underline{1}, \underline{1}, \mathbf{2}) \longleftrightarrow$$

$$\longleftrightarrow (0, 0, 1, 1, 0, 0) \longleftrightarrow (0, 0, 0, 0, \underline{1}, \mathbf{1}) \longleftrightarrow$$

$$\longleftrightarrow (0, 0, 0, 0, 0).$$

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):

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 **NP**[1] – it is known neither whether it is **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 **NP**-complete.

It is difficult to answer precisely the question at the beginning of this subsection. 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 6.1.16 on page 389). 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.

---

[1]Wikipedia, *NP (complexity)*, http://en.wikipedia.org/wiki/NP_(complexity) (as of Aug. 7, 2013, 13:44 GMT).

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:

$$K_{1,1001},\ K_{7,143},\ K_{11,91} \text{ and } K_{13,77}. \qquad \square$$

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) $5 \cdot 4 \cdot 3 = 60$. $\qquad \square$

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:

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 $V = (v_1, \ldots, v_n)$ is the sequence

$$(\deg v_1, \deg v_2, \ldots, \deg v_n).$$

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

$$(1) \qquad \sum_{v \in V} \deg v = 2|E|.$$

In particular, the sum of the degree sequence must be even.

The following theorem[2] of Havel and Hakimi is a first result about operations with graphs. 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.

Theorem. *For any natural numbers $0 \le d_1 \le \cdots \le d_n$, 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*

$$(d_1, d_2, \ldots, d_{n-d_n} - 1, d_{n-d_n+1} - 1, \ldots, d_{n-1} - 1).$$

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.

---

[2]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

Solution. The adjacency matrix of the given graph is

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

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. □

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:
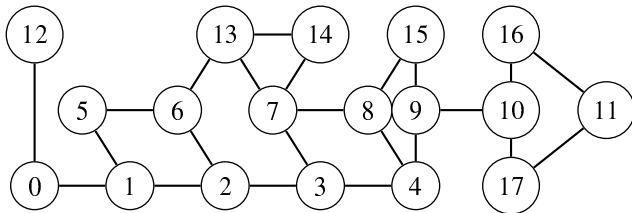
13.A.24. Prove that a Hamiltonian graph (see 13.1.13) must be 2-vertex-connected. Give an example of a graph which is 2-vertex-connected yet does not contain a Hamiltonian cycle.

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). □

The reverse implication is more difficult. The following needs to be proved. Suppose a fixed degree sequence $(d_1, \ldots, d_n)$ with $0 \leq d_1 \leq \cdots \leq d_n$ is given. Then there exists a graph whose vertex v_n is adjacent to exactly the last d_n vertices $v_{n-d_n}, \ldots, v_{n-1}$.

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. The idea is 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_ℓ 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 procedure reveals that the degree sequence of a graph falls far short of determining the graph.

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:

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 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

The *adjacency matrix* of the (undirected) graph $G = (V, E)$ is defined as follows. Fix a (total) ordering of its vertices $V = (v_1, \ldots, v_n)$. 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 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, and 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 $V = (v_1, \ldots, v_n)$, 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_ℓ to a walk of length k between v_ℓ 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_ℓ 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 $a_{ij}^{(k+1)}$. □

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).*

also 5! paths of length six. Clearly, there are no longer paths in K_7. Altogether, we have $1+5+5\cdot4+5\cdot4\cdot3+5!+5! = 326$ 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 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 $C = v_1v_2 \ldots v_{3n}$. 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 $v_1v_2 \ldots v_iv_jv_{j+1} \ldots v_{3n}$ and $v_iv_{i+1} \ldots v_j$ 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, $1 \le i < j \le 3n$, which are connected to the same vertex outside the cycle.
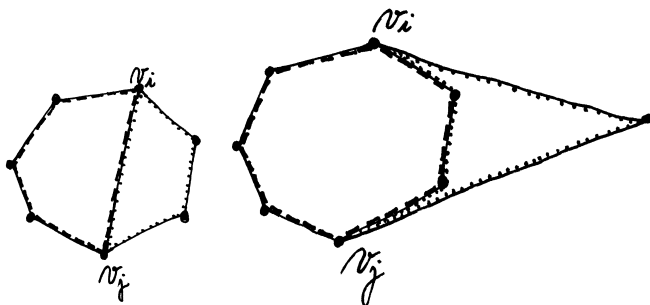
PROOF.
$(A + \mathbb{E}_n)^{n-1} = A^{n-1} + \binom{n-1}{1}A^{n-2} + \cdots + \binom{n-1}{n-2}A + \mathbb{E}_n.$

The entries of the resulting matrix are (using the notation as above)

$$a_{ij}^{(n-1)} + \cdots + \binom{n-1}{\ell}a_{ij}^{(n-1-\ell)} + \cdots + (n-1)a_{ij} + \delta_{ij},$$

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 $0, \ldots, n - 1$ 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. □

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

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
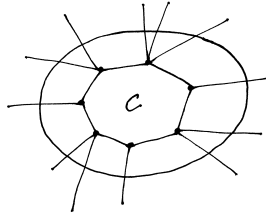
can consider the graph which is obtained from the original one by replacing the vertices $v_1, v_2, \ldots, v_{3n}$ with a single vertex V.

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 $w_1 w_2 \ldots w_k$ 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. □

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 $1, 2, \ldots, 6$. 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 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

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. An (undirected) 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*[3] 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 . 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.*

---

[3]Karl Menger proved this as early as in 1927; that is, before graph theory came into being.

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 $(3, 6)$. $\qquad\square$

13.B.2. Consider a graph on six vertices which are labeled $1, 2, \ldots, 6$. 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.) $\qquad\square$

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 $1, 2, \ldots, 6$. Write the order of edges of K_6 in which they are visited by the depth-first search algorithm, supposing the search is initiated from vertex 3 and the neighbors of a given vertex are visited in ascending order. $\qquad\bigcirc$

13.B.4. Let the vertices of the graph K_6 be labeled $1, 2, \ldots, 6$. Write the order of edges of K_6 in which they are visited by the breadth-first search algorithm, supposing the search is

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:

$$(v = v_0, e_1, v_1, e_2, \ldots, v_{k+1} = w)$$

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 cosequence 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

initiated from vertex 3 and the neighbors of a given vertex are visited in ascending order.

13.B.5. Apply Dijkstra' algorithm to find the shortest path from vertex number 9 to each of the other vertices.
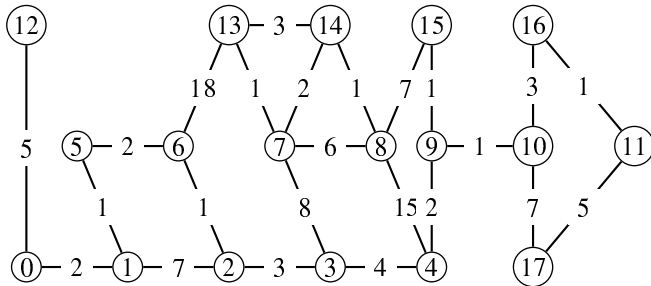
13.B.6. Give an example of

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):
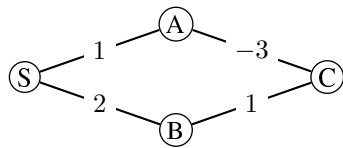
If Dijkstra's algorithm is run from S, then it visits the vertex A and fixes its distance from S to 1. However, there is a shorter path, namely the path (S, B, C, A) of length 0. As for the second problem, consider the following:
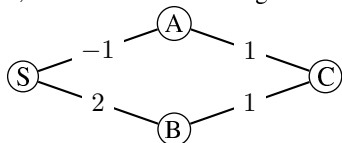
Bellman-Ford algorithm. This algorithm is based on the same principle as Dijkstra's. However, instead of going through particular vertices, it processes them "simultaneously"–the *relaxation* loop (i. e., finding out whether the temporary distances of the vertices can be improved using a given edge) is iterated $(|V|-1)$-times over all edges. The advantage

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. □

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 a 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.
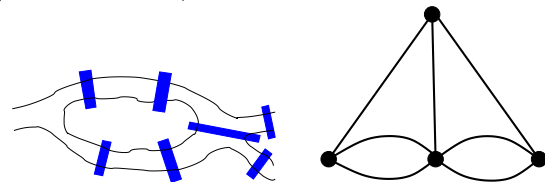
The situation is depicted in the diagram. On the left, there is a sketch of the river with the islands and bridges. The corresponding multigraph is caught in the right-hand diagram. The vertices of this graph correspond to the "connected land", while the edges correspond to the bridges. If it is desired 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 $(v_0, e_1, \ldots, v_k)$ in G

is that this approach works even with negative edges, and it 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	∞	∞	∞	1	∞	22	3(6)	4
2	0	∞	23	∞	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	∞	∞	∞	1	∞	-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. □

Paths between all pairs of vertices. We often need to know the shortest paths between all pairs of vertices. Of course, we

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 goes along the edge e – a contradiction. □

Corollary. *A 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 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 finding an 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.

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 $O(n^3)$;

- 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 $U_0, U_1, \ldots, U_{|V|}$, 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 $\{1, 2, \ldots, k\}$;

- 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 $1, \ldots, k$, 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 $1, \ldots, k-1$). Otherwise, the wanted path is also the shortest path from i to j which can go only through the vertices $1, \ldots, k-1$. 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 $v, w = p_{uv}, p_{uw}, \ldots$.

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

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 $O(1{,}567^n)$.[4]

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 $P = (v_0, \ldots, v_k)$ 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.

---

[4]Björklund, Andreas (2010), "Determinant sums for undirected Hamiltonicity", Proc. 51st Impartial Symposium on Foundations of Computer Science (FOCS '10), pp. 173-182, arXiv:1008.0541, doi:10.1109/FOCS.2010.24.

Solution. We proceed according to the algorithm, obtaining the following shortest-length matrices and predecessor matrices:

$$U_0 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \infty & 10 & 0 & 3 \\ 5 & 6 & 6 & 0 \end{pmatrix}, \quad P_0 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ - & 3 & - & 3 \\ 4 & 4 & 4 & - \end{pmatrix};$$

$$U_1 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \infty & 10 & 0 & 3 \\ 5 & 6 & \mathbf{2} & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ - & 3 & - & 3 \\ 4 & 4 & 1 & - \end{pmatrix};$$

$$U_2 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \mathbf{7} & 10 & 0 & 3 \\ \mathbf{3} & 6 & -1 & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ 2 & 3 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix};$$

$$U_3 = \begin{pmatrix} 0 & 4 & -3 & \mathbf{0} \\ -3 & 0 & -7 & \mathbf{-4} \\ 7 & 10 & 0 & 3 \\ 3 & 6 & -1 & 0 \end{pmatrix}, \quad P_3 = \begin{pmatrix} - & 1 & 1 & 3 \\ 2 & - & 2 & 3 \\ 2 & 3 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix};$$

$$U_4 = \begin{pmatrix} 0 & 4 & -3 & 0 \\ -3 & 0 & -7 & -4 \\ \mathbf{6} & \mathbf{9} & 0 & 3 \\ 3 & 6 & -1 & 0 \end{pmatrix}, \quad P_4 = \begin{pmatrix} - & 1 & 1 & 3 \\ 2 & - & 2 & 3 \\ 2 & 4 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix}.$$

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–ChvÄ Ątal 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

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 (eg. 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

not been adjacent and $\deg(u) + \deg(v) \geq n$ until no such pair of vertices u, v exists.

Bondy, ChvĂĄtal: A graph G is Hamiltonian if and only if $cl(G)$ is.

13.B.9. Prove the Bondy–ChvĂĄtal 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–ChvĂĄtal 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.

i) If a graph G satisfies Ore's condition, then its closure is a complete graph, which is Hamiltonian, of course. By the Bondy–ChvĂĄtal 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:
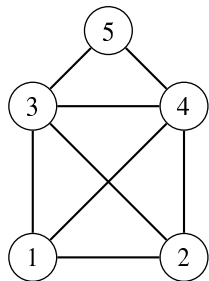
The degree of vertex 5 is 2, which is less than $\frac{5}{2}$. The sum of the degrees of any pair of (not only non-adjacent) vertices is at least 5.

iii) The wanted conditions are satisfied by the cycle graphs $C_n, n > 4$, for which $cl(C_n) = C_n$. □

Planar graphs.

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 $n = 2^k - 1$, $k = 1, 2, \ldots$. 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 inreasing (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!
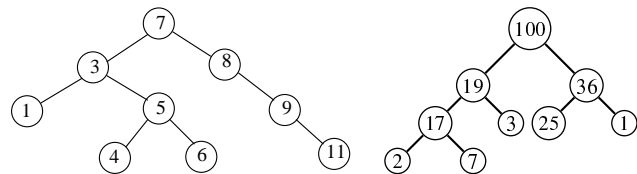
The left-hand diagram shows a binary search tree. In the right-hand diagram, there is a max heap.

Much literature is devoted to trees, their applications and miscellaneous variations.

13.B.11. Decide whether the graph in the picture is planar.
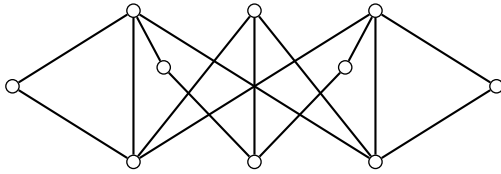
Solution. By the Kuratowski theorem (see page 945), this graph is not planar since one of its subgraphs is a subdivision of $K_{3,3}$. □

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? ○

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 lies belongs to two faces. If f is the number 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.

□

13.B.14. Decide whether the given planar graph is maximal. Add as many edges as possible while keeping the graph planar.
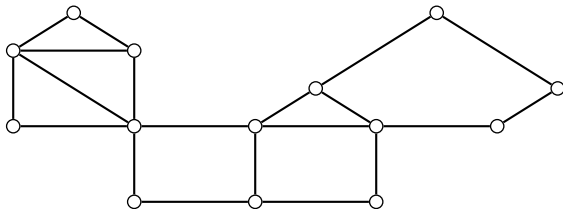
Solution. The graph has 14 vertices and 20 edges, hence $3|V| - 6 - |E| = 16$. Therefore, it is not maximal, and 16 edges can be added so that it is still planar.
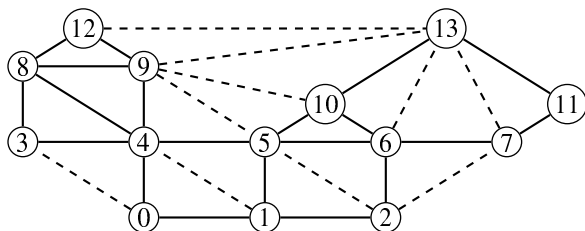
13.1.17. Remarks on sorting. Suppose it is required to distinguish all different sortings of n elements, thus distinguishing among $n!$ different objects. If there is no information other than comparing the order of two single elements, then the tree of all possible decision paths can be written down. 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 $h \log 2 > \log n!$.

$$\log n! = \log 1 + \log 2 + \cdots + \log n$$
$$> \int_1^n \log x \, dx = n \log n - (n-1)$$
$$h > \frac{n \log n - n}{\log 2} > n \log n - n$$

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 occurrencies 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 ν 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 $\varphi : T \to T'$ 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 ν and ν'.

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.
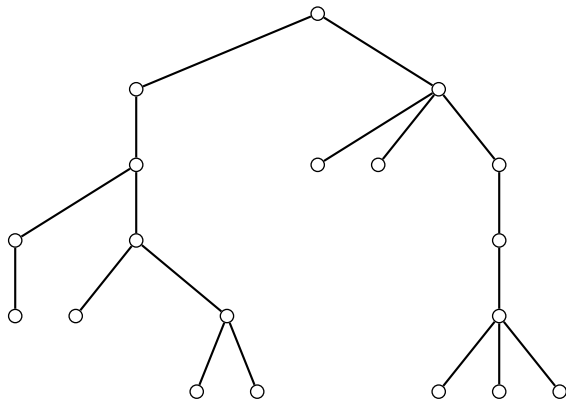
Solution.

i) Using the procedure from 13.1.18, we get the following code of the plane tree:

0 0 0001100100101111 1 0 01010000101011111 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.

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 $W = 0W_1 \ldots W_\ell 1$, where W_i are the ℓ 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 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):

$$0W_3W_81 \mapsto 00W_1W_510W_911$$
$$\mapsto 00010W_4W_61100W_{11}111$$
$$\mapsto 000100101110001111.$$

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 $W = W_1W_2 \ldots W_\ell$. 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. □

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.

□

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,
- 0000011001001011111001010001010111111.

◯

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*.

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).

The same construction can be used as for the plane trees, ordering the vertices lexicographically with respect to their codes (see 12.5.5 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 $W_1, \ldots, W_\ell$, then the code of the vertex v is

$$0W_1 \ldots W_\ell 1,$$

where the order is selected so that $W_1 \leq W_2 \leq \cdots \leq W_\ell$.

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 $\mathrm{ex}_T(v)$ 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 maximal 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 tree isomorphism can be implemented in linear time with respect to the number of vertices of the trees.

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.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.
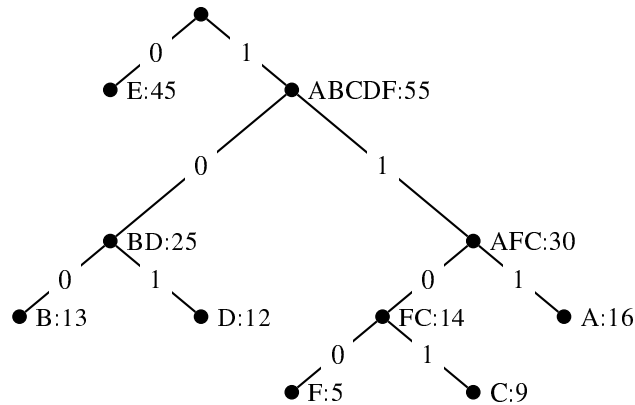
We have thus obtained the codes A : 111, B : 100, C : 1101, D : 101, E : 0, F : 1100. Multiplying the code lengths by the frequencies, we can see that a 100-letter message with the given distribution of letters is encoded into only

$$3 \cdot 16 + 3 \cdot 13 + 4 \cdot 9 + 3 \cdot 12 + 1 \cdot 45 + 4 \cdot 5 = 224$$

bits. □

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). □

13.C.2. Let the vertices of K_6 be labeled $1, 2, \ldots, 6$ 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

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 $c : [0, 1] \to \mathbb{R}^2$ 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:
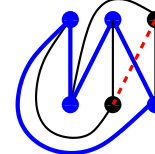
For a complete proof, more mathematical tools are needed. A complete explanation is not provided here, but an indication of the reasoning follows.

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).

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

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.
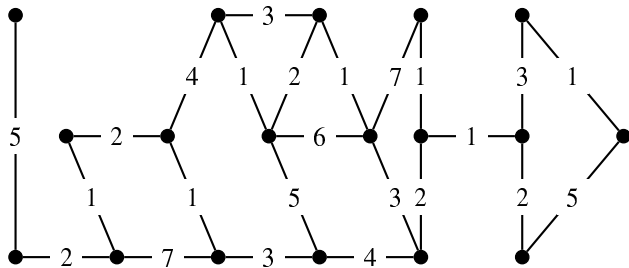
Solution. The spanning tree is
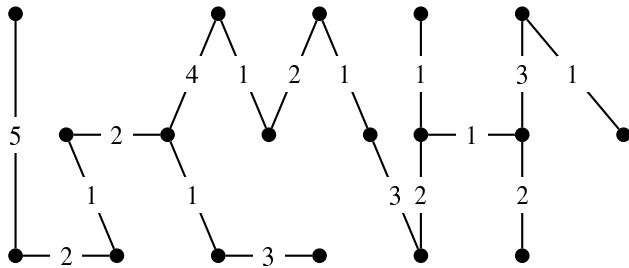
Borůvka's algorithm cannot be applied directly since the mapping of the weights to the edges is not injective. However, this can be fixed easily by slight modifications of the weights. $\square$

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. $\bigcirc$

13.C.5. We are given the following table of distances of world metropolises: London, Mexico City, New York, Paris,

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 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, staring 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 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

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 JarnĂŋk-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:
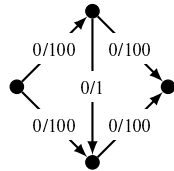
The bad behavior of the algorithm is due to the fact that it uses depth-first search to find unsaturated paths.

Solution. We proceed strictly by depth-first search (examining the vertices from left to right and then top down):
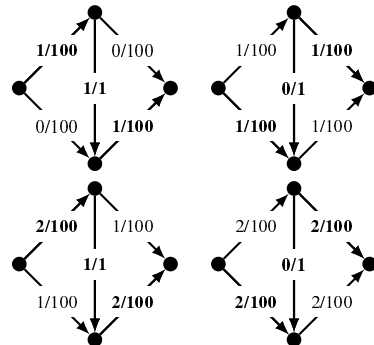
number of faces is independent of the particular embedding of the graph in the plane:

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. □

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 $G_1, \ldots, G_k$, $k \geq 2$. 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 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 $3|S| = 2|E|$.
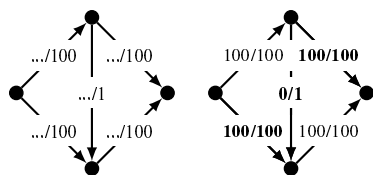
We can see that 200 iterations were needed in order to find the maximum flow. □

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). □

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. □

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. □

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:

947

13.D.4. Find a maximum flow and a minimum cut in the following flow network (source=1,sink=14).
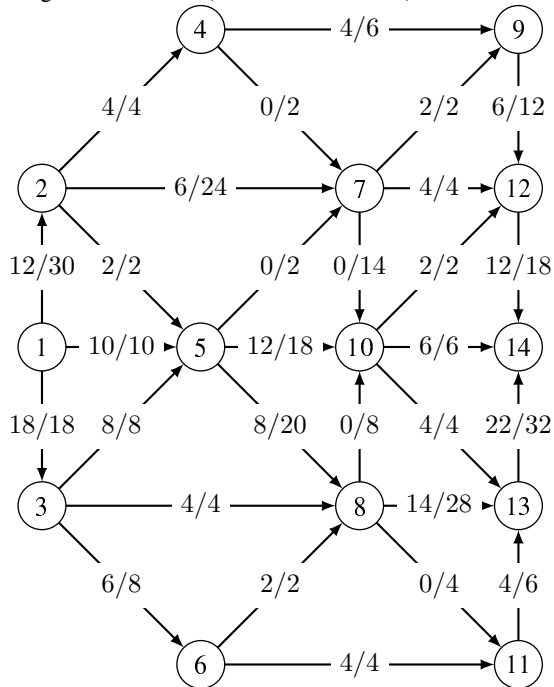
Solution. The paths are saturated in the following order:

$$1 \xrightarrow{18} 2 \xrightarrow{18} 7 \xrightarrow{14} 10 \xleftarrow{12} 5 \xrightarrow{12} 8 \xrightarrow{4} 11 \xrightarrow{2} 13 \xrightarrow{10} 14 \qquad r.2$$

$$1 \xrightarrow{16} 2 \xrightarrow{16} 7 \xrightarrow{12} 10 \xleftarrow{10} 5 \xrightarrow{10} 8 \xrightarrow{14} 13 \xrightarrow{8} 14 \qquad r.8$$

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

$[2,4] : 4, [7,9] : 2, [7,12] : 4, [10,12] : 2, [10,14] : 6, [13,14] : 32.$ □

13.D.5. Find a maximum flow in the following network on the vertex set $\{1, 2, \ldots, 9\}$ 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:
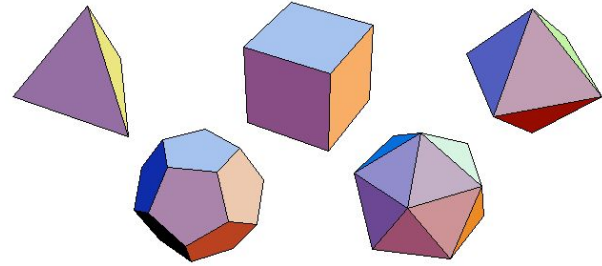
Translate the condition of regularity to the properties of the corresponding graphs: Every vertex needs the same degree $d \geq 3$, and the boundary of every face must contain the same number $k \geq 3$ of vertices. Let n, e, s denote the total number of vertices, edges, and faces, respectively.

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

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:
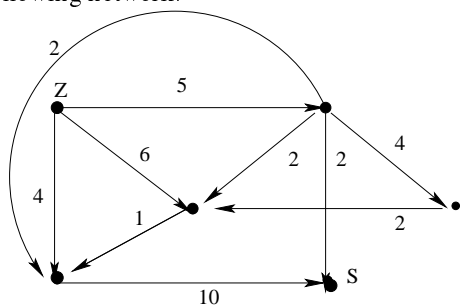
Solution. Let us fix the following edge labeling:
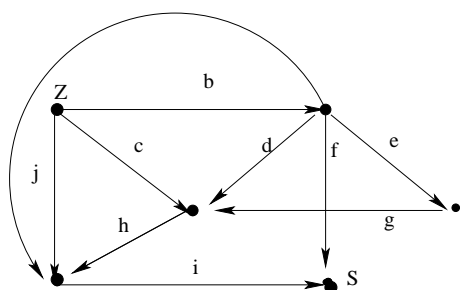
Then, there are cuts: {f,i},{f,h,j,a},{f,j,c,a,d,e},{f,j,c,a,d,g}, {b,j,c},{b,j,h},{b,i}. Their capacities are 12, 9, 20, 18, 15, 10, and 15, respectively. $\square$

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. Algorithms and 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*.

13.D.7. Find a maximum flow in the network given in the above exercise. ◯

Further exercises on maximum flows and minimum cuts can be found on page .

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 $1/2^n$. □

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

$$P \left(\bigcup_{i=1}^{n} A_i \right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n))$$

for every $n \in \mathbb{N}$. We are looking for an $n \in \mathbb{N}$ such that

$$P \left(\bigcup_{i=1}^{n} A_i \right) =$$
$$1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n)) > 0.9.$$

Apparently, we have $P(A_i) = 1/10$ for any $i \in \mathbb{N}$. Therefore, it suffices to solve the inequality

$$1 - \left(\tfrac{9}{10} \right)^n > 0.9,$$

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. □

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.

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 $\mathrm{O}(1)$ (constant time), while the others are in $\mathrm{O}(n)$ (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 situations:

- *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*.

Solution. By the theorem on product of probabilities, the result is

$$\frac{28}{32} \cdot \frac{27}{31} \cdot \frac{26}{30} \cdot \frac{25}{29} \cdot \frac{24}{28} \cdot \frac{4}{27} \doteq 0.0723. \qquad \square$$

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, 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

$$\frac{4}{32} \cdot \frac{3}{31} + \frac{28}{32} \cdot \frac{4}{31} = \frac{1}{8}. \qquad \square$$

13.E.5. **Combinatorial identities.** Use combinatorial means to derive the following important identities (in particular, do not use induction):

Arithmetic series $\quad \sum_{k=0}^{n} k = \frac{n(n+1)}{2} = \binom{n+1}{2}$

Geometric series $\quad \sum_{k=0}^{n} x^k = \frac{x^{n+1}-1}{x-1}$

Binomial theorem $\quad (x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$

Upper binomial theorem $\quad \sum_{k=0}^{n} \binom{k}{m} = \binom{n+1}{m+1}$

Vandermonde's convolution[1] $\quad \binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}$.

○

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, $) \leq k \leq 20$ 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

$$P(k\, redballs) = \binom{20}{k}(0.3)^k(0.7)^{20-k}$$

$\square$

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, $) \leq k \leq 20$ if the sampling is done without replacement (repetition not allowed)?

---

[1] Also called hockey identity.

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:
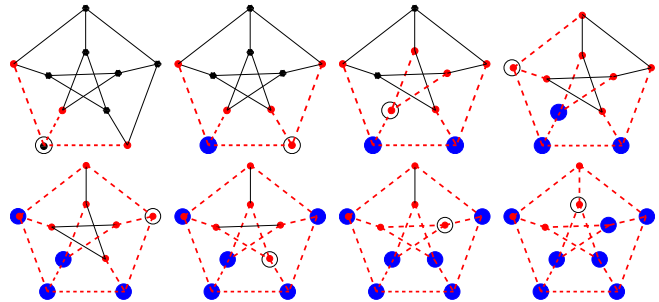
The first 8 steps are shown here. The circled vertex is the one to be processed, the bold vertices are the already processed one, while the dashed edges are those that have been processed, and the small vertices adjacent to some dashed edges are the active ones. At the given vertex, the edges are processed 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.
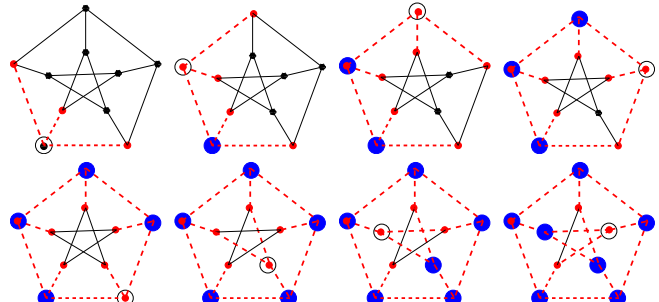
As a simple example of graph searching, consider an algorithm for finding all connected components of a given graph. The only information that must be processed during the search (no matter whether breadth-first or depth-first) is which component is being examined.

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.

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 $|S| \binom{(}{1}00, 20)$. 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,

$$|A| = \binom{(}{3}0, k)\binom{(}{7}0, 20 - k).$$

Thus,

$$P(A) = \frac{|A|}{|S|} = \frac{\binom{(}{3}0, k)\binom{(}{7}0, 20 - k)}{\binom{(}{1}00, 20)}.$$

$\square$

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 $\frac{1}{2}$.

 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

$$P(A_k) = 1 - \frac{P_k^12}{/}12^k, \text{for} 2 \leq k \leq 12.$$

For $k > 12$, obviously, $P(A_k) = 1$. Therefore the required

$$P = \frac{1}{4}P(A_5) + \frac{1}{4}P(A_10) + \frac{1}{2}P(A_{15}) = \frac{1}{4}(1 - \frac{P_5^12}{/}12^5) + \frac{1}{4}(1 - \ldots$$

The second part of the question asks for conditional probability $P(k = 10|A)$. According to Bayes' rule:

$$P(k = 10|A) = \frac{P(A|k = 10)P(k = 10)}{P(A)} = \frac{P(A_{10})}{4P(A))} = \frac{(1 - \frac{P_{10}^12}{/}12^10)}{(1 - \frac{P_5^12}{} 12^5) + (1 - \frac{P_5}{} 12^{10}) + 2}$$

$\square$

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?

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 $d_G : V \times V \to \mathbb{N}$ 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 482):

- $d_G(v, w) \geq 0$, 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 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 consecutively 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*[5].

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

_____

[5]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

Solution. Let A_i be the 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:

$$P(E) = P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i) - \sum_{i<j} P(A_i \cap A_j) +$$

$$\sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{N-1} P(\cap_{i=1}^{N} A_i)$$

Note, that due to symmetry, the probability of any such intersection depends only on the number of sets A_i in the intersection, so

$$\sum_{i=1}^{N} P(A_i) = NP(A_1);$$

$$\sum_{i<j} P(A_i \cap A_j) = \binom{(}{N}, 2)P(A_1 \cap A_2);$$

$$\sum_{i<j<k} P(A_i \cap A_j \cap A_k) = \binom{(}{N}, 3)P(A_1 \cap A_2 \cap A_3);$$

$$\ldots$$

Therefore,

$$P(E) = NP(A_1) - \binom{(}{N}, 2)P(A_1 \cap A_2) + \binom{(}{N}, 3)P(A_1 \cap A_2 \cap A_3) - \ldots + (-1)^N P(\cap_{i=1}^{N} A_i).$$

We need to find $P(A_1), P(A_1 \cap A_2), P(A_1 \cap A_2 \cap A_3), \ldots$ 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 $|A| = (N-1)!$. Similarly

$$P(A_1 \cap A_2) = \frac{|A_1 \cap A_2|}{N!} = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)};$$

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

Thus,

$$P(E) = N\frac{1}{N} - \binom{(}{N}, 2)\frac{1}{N(N-1)} + \binom{(}{N}, 3)\frac{1}{N(N-1)(N-2)} - \ldots + (-1)^N \frac{1}{N!}$$

$$1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^N \frac{1}{N!},$$

which approaches $1 - \frac{1}{e}$ as $N \to \infty$. $\square$

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?

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 ∞, 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).

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 $\binom{13\cdot4}{2} = 1326$. 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 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

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,

$$(1) \qquad d_w(v) \leq d(v)$$

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 $0 = d_0 < \cdots < d_k$ 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 δ 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 $(v_0, e_1, v_1, \ldots, v_\ell, e_{\ell+1}, y)$ with total length d_{j+1}. However, then

$$(2) \qquad d_w(v_\ell) \leq d_{j+1} - w(\{v_\ell, y\}) < d_{j+1}.$$

It follows from the assumption that the vertex v_ℓ 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.

$\binom{52-5}{2}$ 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 from 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 $0 + 3 + 4 + \cdots + 4 + 2 = 45$ 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 %. □

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. ○

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 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. □

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 ne needed in the implementation. However, there is a better solution than that. For instance, if the shortest path in a road network is required, the known air distances can be used as the initialization values. Then, the bounds for the distances $d_w^0(v)$ between vertices v and v_0 can be used such that for any edge $e = \{v, y\}$,

$$|d_w^0(v) - d_w^0(y)| \le w(e).$$

This is sufficient for the proof of the correctness of the algorithm. (Check this yourself!)

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.

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:
$$2n^n - n!. \qquad \square$$

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 $P\left(H^C\right)$ of the complementary event to H is 2^{-5}. Further, the probability of A is the same as the probability $P\left(A^C\right)$. 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
$$P(A|H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{1}{2}}{1 - \left(\frac{1}{2}\right)^5} = \frac{16}{31}. \qquad \square$$

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.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

Input: Graph $G = (V, E)$
Output: A forest $T = (V, E')$ consisting of spanning trees of the components of G.
(1) Sort all edges $e_0, \ldots, e_m \in E$ 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 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.
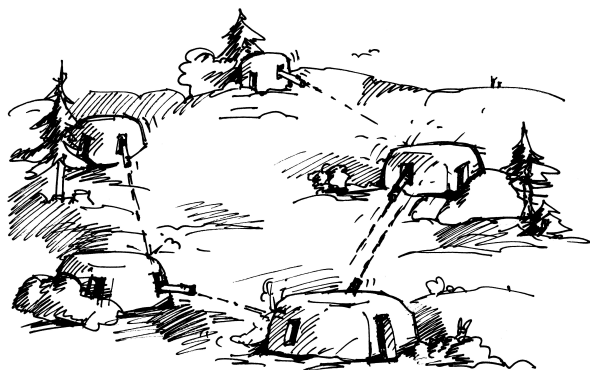
Solution. First of all, note that a set of hit fortresses is a possible result if and only if no pair of adjacent-but-one fortresses (i. e., whose numbers differ by 2 modulo n) is unhit. Therefore, if n is odd, then $P(n)$ is equal to the number $K(n)$ of results where no pair of *adjacent* fortresses is unhit (consider the order $1, 3, 5, \ldots, n, 2, 4, \ldots, n-1$). 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)

$$(K(n), K(n-1)) = (K(n) - K(n-1), K(n-1)) =$$
$$= (K(n-2), K(n-1)) = \cdots = 1.$$

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

$$(K(a), K(2a+1)) = (K(a), F(2)K(2a) + F(1)K(2a-1))$$
$$= (K(a), F(3)K(2a-1) + F(2)K(2a-2)) = \ldots$$
$$= (K(a), F(a+1)K(a+1) + F(a)K(a))$$
$$= (K(a), F(a+1)) = (F(a+2) - F(a-2), F(a+1))$$

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 it 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 $O(n+m)$. □

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[6]:

---

[6]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

$$= \ (F(a+2) - F(a+1) - F(a-2), F(a+1))$$

$$= \ (F(a) - F(a-2), F(a+1))$$

$$= \ (F(a-1), F(a+1)) = (F(a-1), F(a)) = 1.$$

$$(K(a), K(2a-1))$$

$$= \ (K(a), F(2)K(2a-2) + F(1)K(2a-3))$$

$$= \ (K(a), F(3)K(2a-3) + F(2)K(2a-4))$$

$$= \ \cdots = (K(a), F(a)K(a) + F(a-1)K(a-1))$$

$$= \ (K(a), F(a-1)) = (F(a+2) - F(a-2), F(a-1))$$

$$= \ (F(a+2) - F(a), F(a-1))$$

$$= \ (F(a+2) - F(a+1), F(a-1)) = (F(a), F(a-1)).$$

This proves the proposition. □

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

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 $\sum_{e \in E(\tilde{T})} w(e) \le \sum_{e \in E(T)} w(e)$, so the goal is to show that also $\sum_{e \in E(T)} w(e) \le \sum_{e \in E(\tilde{T})} w(e)$.

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 $ee_1e_2 \cdots e_k$ 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 $w(e_i) \ge w(e)$.

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. □

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*.
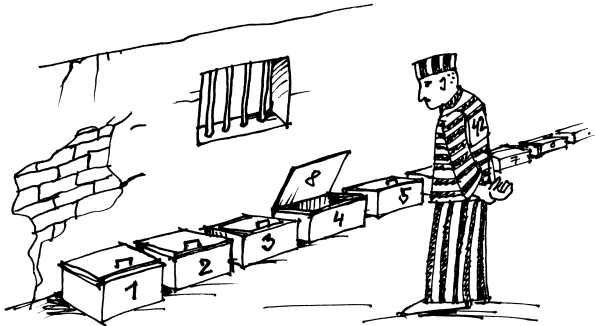
Probability of success. The guard's possible placements of the balls bijectively correspond to permutations of the numbers 1 through 100. In order to find the probability of success, we must realize for which permutations the pointer strategy works. Recall that every permutation can be expressed as the 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

$$\binom{n}{r}(r-1)!\,(n-r)! = \frac{n!}{r}$$

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:

Jarník-Prim's algorithm[7]

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

---

[7]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.

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 $\sum_{k=1}^{n} \frac{1}{k} \to \ln(n) + \gamma$ for $n \to \infty$, where γ is Euler's constant. Thus, we have:

$$p = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=1}^{\frac{n}{2}} \frac{1}{k} \to \ln(n) + \gamma - \ln\left(\frac{n}{2}\right) - \gamma = \ln 2, \text{ pro } n \to \infty.$$

Hence it follows that, for large values of n, the probability of success approaches $1 - p \simeq 1 - \ln 2 = 0.30685\ldots$. 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

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 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.

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:

$$(7, 5, 8)(2, 4)(1, 6, 3) \rightarrow (6, 3, 1)(4, 2)(8, 7, 5) \rightarrow$$
$$\rightarrow (6, 3, 1, 4, 2, 8, 7, 5).$$

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 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

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, 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.
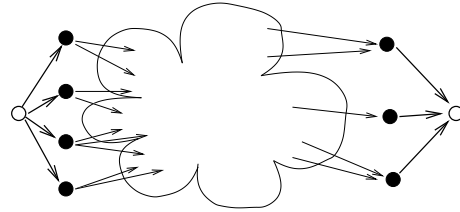
FLOW NETWORKS

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).
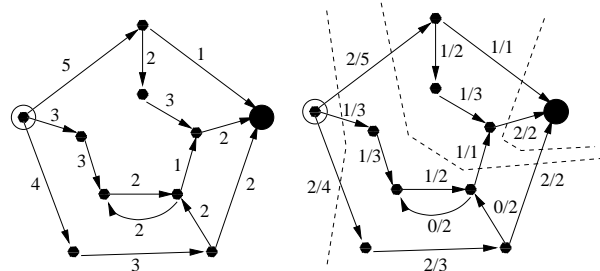
probability of success. Therefore, the pointer strategy is better than or equally good as any other strategy. □

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 ({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:

$$\frac{x(x-1)}{2} + \frac{(n-x)(n-x-1)}{2} = \frac{2x^2 - 2nx + n^2 - n}{2} =$$
$$= \left(x - \frac{n}{2}\right)^2 + \frac{n^2}{4} - \frac{n}{2} \geq \frac{n^2}{4} - \frac{n}{2} = \frac{(n-1)^2}{4} - \frac{1}{4}.$$

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 $P(X > k) > 0$.

Consider the random variables X_i for $i = 1, 2, \ldots, m$ 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:

$$X = X_1 + X_2 + \cdots + X_m$$

Using the linearity of expectation, we obtain:

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_m].$$

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 are added up through which each path from z to s must go. 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 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*[8], 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.

---

[8]Ford, L. R.; Fulkerson, D. R. (1956). "Maximal flow through a network". Canadian Journal of Mathematics 8: 399–404.

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-the 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-the 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

$$(n - 2t_i)^2 \geq 1 \iff t_i \leq \frac{n-1}{2} \text{ or } t_i \geq \frac{n+1}{2},$$

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:

$$E[X] \geq m \frac{\left(\frac{n-1}{2}\right)^2}{\frac{n(n-1)}{2}} = \frac{m(n-1)}{2n}.$$

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.)

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

---

Input: A network $S = (V, E, z, s, w)$.
Output: A maximal possible flow $f : E \to \mathbb{R}$ and a minimal cut C, which is given by those edges which lead from U 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).

For the same reason, all edges e leading from $V \setminus U$ to U must have $f(e) = 0$.

Clearly, 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

$$|C| = \sum_{\text{edges from } U \text{ to } V \setminus U} f(e) - \sum_{\text{edges from } V \setminus U \text{ to } U} f(e),$$

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. $\square$

probability distribution $\mathrm{Bi}(n, x)$ 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 12.2.6). 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. □

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-the round
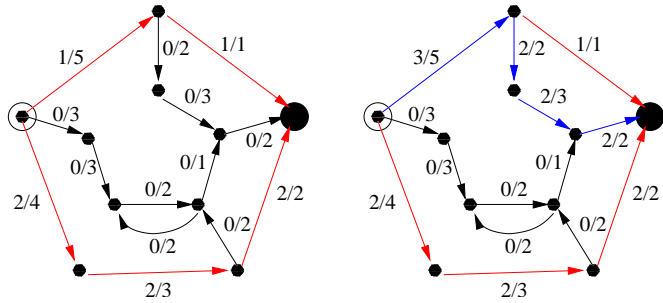
The run of the algorithm is illustrated in two diagrams. On the left, there are two shortest unsaturated paths from the source to the sink in gray (the upper one has two edges, while the lower one has three). On the right, another path is saturated (taking the first turn in the upper path), also drawn 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 $\mathrm{O}(f|E|)$, 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 $\mathrm{O}(|V||E|^2)$.[9] 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

Put some nice example in the other column, e.g. the https://www.cs.princeton.edu/courses/archive/spring13/cos423/lectures/07DemoFordFulkersonPathological.pdf one

_____

[9]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.

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 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 $\{0, 1, \ldots, 2^t - 1\}$. In the i-th round, player α 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 α is indeed α (since $(\alpha \oplus i) \oplus i = \alpha$). Moreover, the second tournament rule is also satisfied: if α plays with β and γ plays with δ in the i-th round (i. e., $\beta = \alpha \oplus i$ and $\delta = \gamma \oplus i$) and if j is such that α plays with γ 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 β indeed plays with δ 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 $|\Gamma| = |\Delta|$ where Γ is the component of A and Δ is the component of B. Actually, we show that any player of Γ competes with a player of Δ in round $i + 1$. Thus, let $C \in \Gamma$, i. e., in G_i, there exists a path $A = X_1, X_2, \ldots, X_m = C$ such that X_j has played with X_{j+1}, $j = 1, \ldots, m-1$, in or before the i-th round. Consider the sequence $Y_1, Y_2, \ldots Y_m$, where Y_k is the opponent of X_k in round $i + 1$, $k = 1, \ldots, m$ (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 $Y_1, \ldots, Y_i$) and in a certain r-the round ($1 \leq r \leq i$), X_j played with X_{j+1} (by the definition of the sequence $X_1, \ldots, X_i$). However, by the second tournament rule, this means that Y_j also played with Y_{j+1} in the r-the round, so the edge $Y_j Y_{j+1}$ is contained in G_i for any $1 \leq j \leq m - 1$, Thus, $Y_1, Y_2, \ldots Y_m$ is a

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.

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.

path in G_i, so $B = Y_1$ and Y_m lie in the same component (Δ). It can be deduced analogously that any player from Δ competes with a player of Γ in round $i + 1$, and since every player plays exactly once in a given round, we must have $|\Gamma| = |\Delta|$. By the definition of a component, the component of A in G_{i+1} is equal to $\Gamma \cup \Delta$. Then, we have either $\Gamma = \Delta$ (then, the component of A in G_{i+1} is Γ), or $\Gamma \cap \Delta = \emptyset$ (in this case, the component of A in G_{i+1} is the disjoint union $\Gamma \cup \Delta$). 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 $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$ of A in the respective graphs $G_1, G_2, \ldots G_k$. We have $|\Gamma_1| = 2$ (since A had exactly one opponent in the first round) and for $1 \le i \le k - 1$, we have either $|\Gamma_i| = |\Gamma_{i+1}|$, or $2|\Gamma_i| = |\Gamma_{i+1}|$. Therefore, the number of vertices (players) of every component is a power of two, i. e., $|\Gamma_k| = 2^l$ for some l, and $\Gamma_k \ge k + 1$ (A had different opponents in the k rounds). Hence, $2^l \ge k + 1$, i. 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. $\qquad\square$

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.

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.

As a simple example, consider a simple variation of the game known as *Nim*.[10]

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 ℓ tokens on the table and the first player to move corresponds to the subtree rooted at F_ℓ. The state with the same number of tokens but the second player to move is represented by the subtree rooted at S_ℓ.
- The vertex F_ℓ 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_ℓ 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 $k \ge 3$, 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:

---

[10]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!".

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.) □

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.

Solution. We are going to prove by induction that any positive integer k satisfies:

$$
\begin{aligned}
g(4k + 1) &= 4k + 1 \\
g(4k + 2) &= 4k + 2 \\
g(4k + 3) &= 4k + 4 \\
g(4k + 4) &= 4k + 3
\end{aligned}
$$

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.

(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_ℓ. 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_ℓ is given the label T. Otherwise, i.e. if all sons are labeled by L, then F_ℓ also gets L.

(3) Similarly, a vertex S_ℓ 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 3×3 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.
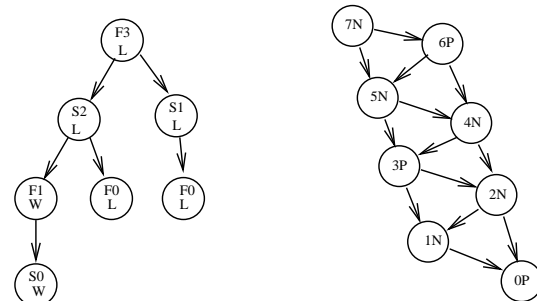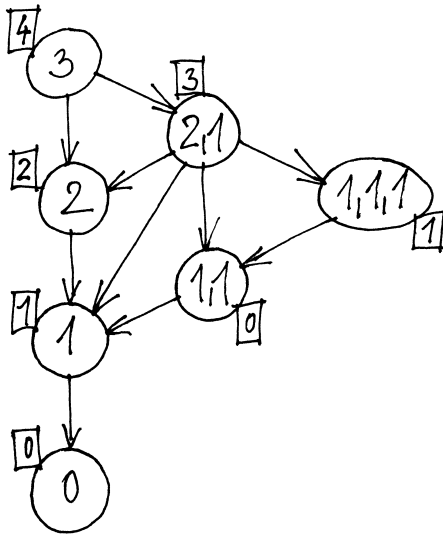
Now, assume that the above is satisfied for all positive integers below $4k + 1$ and let us prove that we indeed have $g(4k + 1) = 4k + 1$. By the definition, the SG-value is the least non-negative integer l such that there is no move to a position with SG-value l. Moreover, this property (including 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 $0, 1, \ldots, 4k$ tokens take all the integers $0, 1, \ldots, 4k$ (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. 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

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.

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 one 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

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

$$(1 + x + x^2 + \ldots)^5 =$$
$$= (1 + x + \ldots)(1 + x + \ldots) \cdots (1 + x + \ldots)$$

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..

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. $g : V \to \mathbb{N}$ on a directed acyclic graph $G = (V, E)$ as follows:[11]

(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 mex is defined on subsets S of the natural numbers $\mathbb{N} = \{0, 1, \ldots\}$ 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

$$(a, b) \mapsto a \oplus b,$$

---

[11] 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.

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.)

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

$$\binom{12+4-1}{4-1} + \binom{11+4-1}{4-1} + \binom{10+4-1}{4-1} + \binom{9+4-1}{4-1}. \quad \square$$

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

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:

<div style="border:1px solid">

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

$$g(v_1 v_2) = g_1(v_1) \oplus g_2(v_2).$$

</div>

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

$$g_1(w_1) \oplus g_2(v_2) = c \oplus g_2(v_2) = g_1(v_1) \oplus b \oplus g_2(v_2)$$
$$= g_1(v_1) \oplus a \oplus g_1(v_1) \oplus g_2(v_2) \oplus g_2(v_2) = a.$$

This fulfills the first part of our plan.

at x^{100} in the product

$$(1 + x + x^2 + \dots)(1 + x^2 + x^4 + \dots)(1 + x^5 + x^{10} + \dots) \cdot$$
$$\cdot (1 + x^{10} + x^{20} + \dots)(1 + x^{20} + x^{40} + \dots) \cdot$$
$$\cdot (1 + x^{50} + x^{100} + \dots) =$$
$$= \frac{1}{1-x} \cdot \frac{1}{1-x^2} \cdot \frac{1}{1-x^5} \cdot \frac{1}{1-x^{10}} \cdot \frac{1}{1-x^{20}} \cdot \frac{1}{1-x^{50}}.$$

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
```

$\square$

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)

$$\frac{x}{x+2} = \frac{x}{2-(-x)} = \frac{x/2}{1-(-x/2)} =$$
$$= \frac{x}{2} - \frac{x^2}{4} + \frac{x^3}{8} - \dots + = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{2^n}.$$

ii) We perform partial fraction decomposition:

$$\frac{x^2+x+1}{2x^3+3x^2+1} = \frac{x^2+x+1}{(x-1)^2(2x+1)} =$$
$$= \frac{A}{2x+1} + \frac{B}{x-1} + \frac{C}{(x-1)^2},$$

finding out that $A = B = \frac{1}{3}$ and $C = 1$; hence

$$\frac{x^2+x+1}{2x^3+3x^2+1} = \frac{1/3}{1+2x} - \frac{1/3}{1-x} + \frac{1}{(1-x)^2} =$$
$$= \sum_{n=0}^{\infty} \left[\frac{1}{3}\left((-2)^n - 1)\right) + (n+1) \right] x^n.$$

$\square$

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) $(1, 1, 2, 2, 4, 4, 8, 8, \dots)$,

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. $\square$

The following useful result is a direct corollary of this theorem:

Corollary. *A vertex $v_1 v_2$ 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.[12] We discuss convex hulls, triangulations, and Voronoi diagrams and focus on a few basic approaches only.

---

[12]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

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.
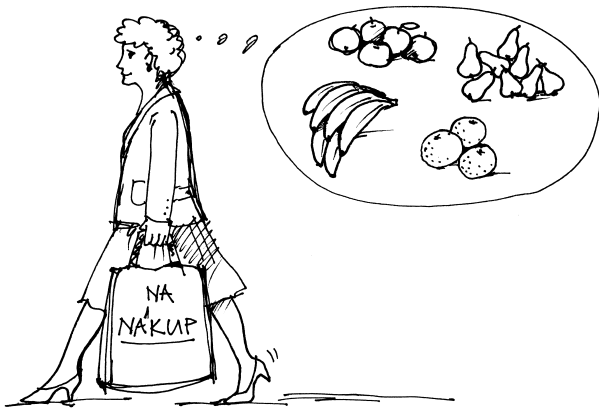
Solution. The generating function for the sequence (a_n) where a_n is the wanted number of ways to buy n pieces of fruit is

$$(1 + x + x^2 + \cdots)(1 + x^2 + x^4 \cdots)(1 + x^4 + x^8 + \cdots) \cdot$$
$$\cdot (1 + x + x^2 + x^3)(1 + x) =$$
$$= \frac{1}{1 - x} \cdot \frac{1}{1 - x^2} \cdot \frac{1}{1 - x^4} \cdot \frac{1 - x^4}{1 - x} \cdot (1 + x) =$$
$$= \frac{1}{(1 - x)^3}.$$

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 $a_n = \binom{n+2}{2}$. □

13.I.6. Using the generalized binomial theorem, prove again the following combinatorial identities:

- $\sum_{k=0}^{n} \binom{n}{k} = 2^n$,
- $\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0$,
- $\sum_{k=0}^{n} k \binom{n}{k} = n2^{n-1}$.

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 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_{active}, until e belonging to $CH(X)$ is found.
  - add e to $CH(X)$ and set its head to be the v_{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.

Solution. Substituting into the binomial theorem

$$(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n$$

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. □

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

$$(1 + x + \cdots + x^{30})(1 + x + \cdots + x^{40})(1 + x + \cdots + x^{50}).$$

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

$$\left(\binom{2}{2} + \binom{3}{2}x + \binom{4}{2}x^2 + \cdots \right)(1 - x^{31} - x^{41} - x^{51} + x^{72} + \ldots).$$

Hence, the coefficient at x^{70} is clearly $\binom{70+2}{2} - \binom{70+2-31}{2} - \binom{70+2-41}{2} - - \binom{70+2-51}{2} = 1061$. □

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. □

13.I.9. Solve the recurrence

$$a_0 = a_1 = 1,$$

$$a_n = a_{n-1} + 2a_{n-2} + (-1)^n.$$

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).[2]

Step 1: $a_n = a_{n-1} + 2a_{n-2} + (-1)^n[n \geq 0] + [n = 1]$.

Step 2: $A(x) = xA(x) + 2x^2A(x) + \frac{1}{1+x} + x$.

---

[2]Despite the statement in *Concrete mathematics*, this sequence can already be found in *The On-Line Encyclopedia of Integer Sequences*.

GIFT WRAPPING
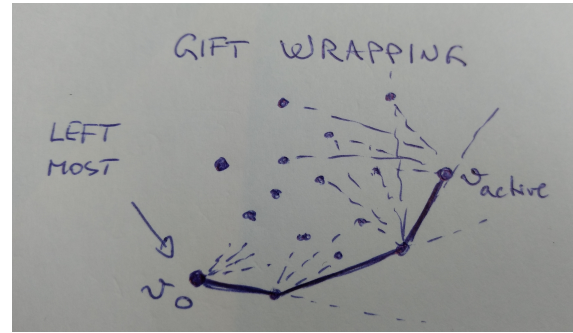
This simple improvement reduces the worst running time of the algorithm to $O(n^2)$. The worst case can be obtained if all the points v_i appear on one circle, and unluckily the right next point is always found as the very last one in partial tests. But the actual running time is much better, at most $O(ns)$, 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 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.

Step 3:

$$A(x) = \frac{1 + x + x^2}{(1 - 2x)(1 + x)^2}.$$

Step 4: $a_n = \frac{7}{9} 2^n + \left(\frac{1}{3} n + \frac{2}{9} \right) (-1)^n.$ $\square$

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 $n - k$.

We thus get the following recurrent formula for the expected number of comparisons:

$$C_n = n - 1 + \sum_{k=1}^{n} \frac{1}{n} \left(C_{k-1} + C_{n-k} \right).$$

It is possible to solve this recurrence (using certain tricks which can be learned to some extent) even without using generating functions.

$$C_n = n - 1 + \frac{2}{n} \sum_{k=1}^{n} C_{k-1} \qquad \text{symmetry of both sums}$$

$$nC_n = n(n-1) + 2 \sum_{k=1}^{n} C_{k-1} \qquad \text{multiply by } n$$

$$(n-1)C_{n-1} = (n-1)(n-2) + 2 \sum_{k=1}^{n-1} C_{k-1} \qquad \text{the same expression for } C_{n-1}$$

$$nC_n = (n+1)C_{n-1} + 2(n-1) \qquad \text{subtracted and rearranged}$$

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.
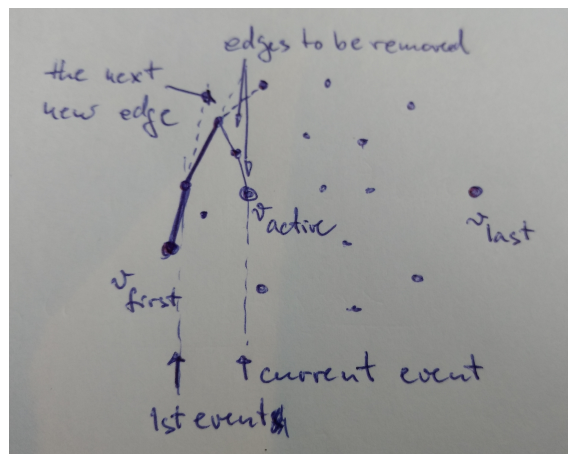
As in the diagram, moving through the events one by one, it is only needed to check, whether the edge joining the last but one active vertex with the current one makes a left or right turn (as usual, right means the clockwise orientation). If right, then add the edge to list, if left, omit the recent edges one by one, until the right turn is obtained.

<div style="text-align:center">SWEEP LINE UPPER CONVEX HULL</div>

Input: A set of points $X = \{v_1, \ldots, v_n\}$ in the plane.
Output: The directed path $UCH(X)$.

(1) *Initialization.*
   - Set the event structure to be the lexicographically ordered list of points $v_{\text{first}}, \ldots, v_{\text{last}}$. 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_{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_{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_{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_{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.

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 $B_n = C_n/n+1$):

$$\frac{C_n}{n+1} = \frac{2(n-1)}{n(n+1)} + \frac{2(n-2)}{(n-1)n} + \cdots + \frac{2 \cdot 1}{2 \cdot 3} + \frac{C_1}{2}.$$

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.I.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 $(F_0, F_2, F_4, \ldots)$. ◯

13.I.12. The *fan* of order n is a graph on $n+1$ vertices, which are labeled $0, 1, \ldots, n$, with the following edges: vertex 0 is connected to all other vertices, and for each k satisfying $1 \le k < n$, 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[3] holds:

$$v_n = v_{n-1} + \sum_{k=0}^{n-1} v_k + 1, \quad v_0 = 0.$$

---

[3] 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.

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 merge them by finding the upper and lower "tangent edge" of those two parts.
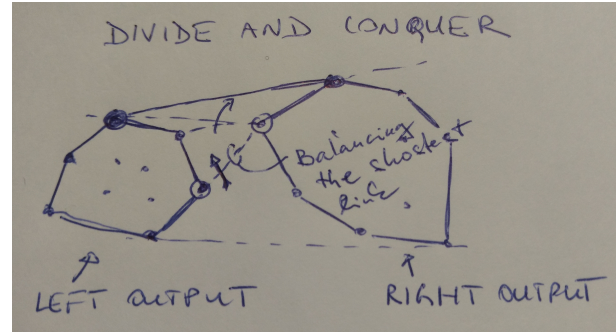
DIVIDE AND CONQUER CONVEX HULL

Input: Set of points $X = \{v_1, \ldots, v_n\}$ in the plane, ordered lexicographically.
Output: The directed path $CH(X)$.

(1) *Divide.* If $n \le 3$, 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 then 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

For a fixed spanning tree of the fan of order n, let k be the greatest integer of the set $\{1, \ldots, n-1\}$ such that the spanning tree contains all edges of the path $(0, 1, 2, 3, \ldots, k)$. This spanning tree cannot contain the edges $\{0, 2\}, \ldots, \{0, k\}, \{k, k+1\}$; therefore, there are the same number of spanning trees for a fixed k as in the fan of order $n - k$ with vertices $0, k+1, k+2, \ldots, n$, 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 $0, 2, \ldots, n$. We have thus obtained the wanted recurrence $v_n = v_{n-1} + v_{n-1} + v_{n-2} + \cdots + v_0 + 1$.

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

$$V(x) = x \cdot V(x) + \sum_{n=0}^{\infty} \sum_{k<n} v_k x^n + \frac{1}{1-x} - 1 =$$

$$= x \cdot V(x) + \sum_{k=0}^{\infty} \sum_{n>k} v_k x^n + \frac{x}{1-x} =$$

$$= \left(\sum_{k=0}^{\infty} v_k x^k \right) \cdot \sum_{n>k} x^{n-k} + \frac{x}{1-x} =$$

$$= \left(\sum_{k=0}^{\infty} v_k x^k \right) \cdot \frac{x}{1-x} + \frac{x}{1-x} = V(x) \cdot \frac{x}{1-x} + \frac{x}{1-x}.$$

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 $v_n = F_{2n}$.
□

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 1×2 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 nor merely convention; there is indeed a unique empty covering).

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 is v_k 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 algorith can now be formulated:

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: $c_n = 2r_{n-1} + c_{n-2} + [n=0]$, $\quad r_n = c_{n-1} + r_{n-2}$.

- Step 2:

$$C(x) = 2xR(x) + x^2 C(x) + 1, \quad R(x) = xC(x) + x^2 R(x).$$

- Step 3:

$$C(x) = \frac{1-x^2}{1-4x^2+x^4}, \qquad R(x) = \frac{x}{1-4x^2+x^4}.$$

- 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, $c_{20} = 413403$. $\qquad \square$

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

$$B(x) = \sum_{b \in B} x^{|b|} = \sum_{n \geq 0} 2^n x^n = \frac{1}{1-2x}.$$

The generating function for the number of ones is

$$C(x) = \sum_{b \in B} j(b) x^{|b|}.$$

Input: A set $X = \{v_1, \ldots, v_n\}$, $n > 3$, 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 uper 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.
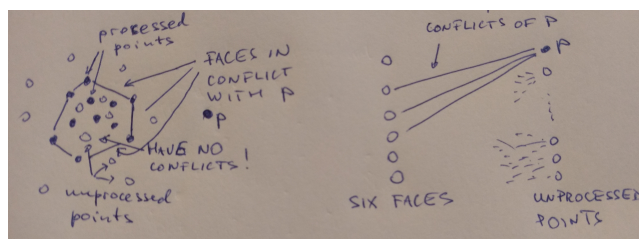
It can be proved that the expected time for this algorithm is $O(n \log n)$, while the worst time is $O(n^2)$. The complete framework for the analysis of randomized algorithm is nicely explained in the book mentioned in the very beginning of this part, see page 972.

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.

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,

$$C(x) = \sum_{b' \in B} (1 + 2j(b'))x^{|b'|+1} = \sum_{b' \in B} x^{|b'|+1} + 2\sum_{b' \in B} j(b')x^{|b'|+1} =$$
$$= xB(x) + 2xC(x).$$

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

$$a_0 = 1, a_1 = 2$$
$$a_n = 4a_{n-1} - 3a_{n-2} + 1 \text{ for } n \geq 2.$$

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

$$a_n = \frac{3}{4}(-1)^k\binom{-1}{n} - \frac{1}{2}(-1)^n\binom{-2}{n} +$$
$$\frac{3}{4}(-3)^n\binom{-1}{n} =$$
$$= \frac{3}{4} - \frac{1}{2}(n+1) + \frac{3}{4}3^n = d\frac{1-2n+3^{n+1}}{4}. \qquad \square$$

13.I.16. Solve the following recurrence using generating functions:

$$a_0 = 1, a_1 = 2$$
$$a_n = 5a_{n-1} - 4a_{n-2} \quad n \geq 2$$

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

$$A(x) = 5xA(x) - 4x^2A(x) - 3x + 1.$$

This is nicely achieved by the *double connected edge lists*.

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 with one of the pointers F_i 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 must be added. Of course the graph of conflicts must also be updated.

Hence
$$A(x) = \frac{1 - 3x}{(1 - 4x)(1 - x)} = \frac{2}{3} \cdot \frac{1}{1 - x} + \frac{1}{3} \cdot \frac{1}{1 - 4x}$$
and
$$a_n = \frac{2}{3}\binom{-1}{n} + \frac{2}{3}\binom{-1}{n}(-4)^n = \frac{4^n + 2}{3}. \qquad \square$$

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

$$G(x) = (1 + x^2 + x^4 + \cdots)(1 + x^5 + x^{10} + \cdots)(1 + x^{10} + x^{20} + \cdots).$$

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

$$[x^{70}]G(x) = -\binom{-3}{7} = \binom{3+7-1}{7} = \frac{9 \cdot 8}{2} = 36. \qquad \square$$

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 tosse sare 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, $\frac{\binom{n}{k}}{2^n}$. $\square$

(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 edges e in R left without one of the 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 disjunct 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 $X = \{v_1, \ldots, v_n\}$ in the plane $\mathbb{R}^2$, there is the obvious equivalence relation $\sim$ on $\mathbb{R}^2$. Define $v \sim w$ if and only if they share the uniquely given closest point $v \in X$. Write $VR(v_i)$ for the equivalence class corresponding to v_i. Define:

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, $\frac{p^k q^{n-k} \binom{()}{n}, k)}{2^n}$. $\qquad\square$

13.I.20. There are coins $C_1, \ldots, C_n$ 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

$$f(x) = \prod_{i=1}^{n} (q_i + p_i x) = \sum_{i=1}^{n} a_m x^m,$$

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,

$$\sum a_{2k+1} = \frac{1}{2}(f(1) - f(-1)) = \frac{1}{2}\left(1 - \frac{1}{2n+1}\right) = \frac{n}{n+1},$$

as $f(1) = 1$ and $f(-1) = \frac{1}{3}\frac{3}{5}\frac{5}{7}\cdots\frac{2n-1}{2n+1} = \frac{1}{2n+1}$. $\quad\square$

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 $P(x) = a_0 + a_1 x + a_2 x^2 + \ldots a_k x^k + \ldots$ is such polynomial. Then its coefficients should satisfy the equation

$$a_0 + 2a_1 + 4a_2 + \ldots 2^k a_k + \ldots = n$$

with integer coefficients $0 \leq a_k \leq 3$. The number of such solutions is given by the coefficient at x^n in

$$f(x) = (1 + x + x^2 + x^3)(1 + x^2 + x^4 + x^6)(1 + x^4 + x^8 + x^{12}) \cdots$$

$$= \prod_{k=0}^{\infty} (1 + x^{2^k} + x^{2(2^k)} + x^{3(2^k)})$$

$$= \prod_{k=0}^{\infty} \frac{1 - x^{2^{k+2}}}{1 - x^{2^k}} = \frac{1}{(1-x)(1-x^2)}$$

$$= \frac{1}{4(1-x)} + \frac{1}{2(1-x)^2} \frac{1}{4(1+x)} = \sum n = 0\infty \left(\frac{1}{4} + (-1) \text{simultaneously} \right) \text{creating}$$

VORONOI DIAGRAM

For a given set of points $X = \{v_1, \ldots, v_n\}$ (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

Hence, the required coefficient is $\lfloor \frac{n}{2} \rfloor + 1$. □

13.I.22. Find the number of subsets of $\{1, \ldots, 2017\}$, such that the sum of its elements is divisible by 5.

Solution. For subset $S = \{s_1, s_2, \ldots, s_k\}$ define $\sigma(S) = x^{s_1} x^{s^2} \ldots x^{s_k}$. Then

$$\sum_{S \subset \{1, \ldots, 2017\}} \sigma(S) = (1+x)(1+x^2) \ldots (1+x^{2017}).$$

The coefficient at x^n in $f(x)$ is the number of subsets in $\{1, \ldots, 2017\}$ with the sum of its elements being exactly n. Let $f(x) = \sum a_n x^n$. We are looking for the sum $A_5 = a_0 + a_5 + a_{10} + \ldots$. Let $\omega_j = \exp(\frac{2j\pi}{5})$, $j = 0, \ldots, 4$ be the 5-th roots of unity with $\omega_0 = 1$. Then

$$A_5 = \frac{1}{5}(f(\omega_0) + f(\omega_1) + f(\omega_2) + f(\omega_3) + f(\omega_4)).$$

Obviuosly $f(\omega_0) = f(1) = 2^{2017}$. For $1 \le j \le 4$

$$f(\omega_j) = ((1+\omega_j^0)(1+\omega_j)(1+\omega_j^2)(1+\omega_j^3)(1+\omega_j^4))^{403}(1+\omega_j)(1+\omega_j^2) = 2^{403}(1+\omega_j+\omega_j^2+\omega_j^3).$$

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}. □

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 $a_0 = 1$. Multiplyuing both sides of the recursion by x^k by summation on k we obtain

$$\sum_{k=0}^{\infty} a_{k+1} x^k = 2 \sum_{k=0}^{\infty} a_k x^k + \sum_{k=0}^{\infty} 4^k x^k.$$

Therefore, for generating function $f(x) = \sum_{k=0}^{\infty} a_k x^k$ 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,

$$f(x) = \frac{1}{1-2x} + \frac{x}{(1-2x)(1-4x)} = \frac{\frac{1}{2}}{1-2x} + \frac{\frac{1}{2}}{1-4x}.$$

$VR(v)$ 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.
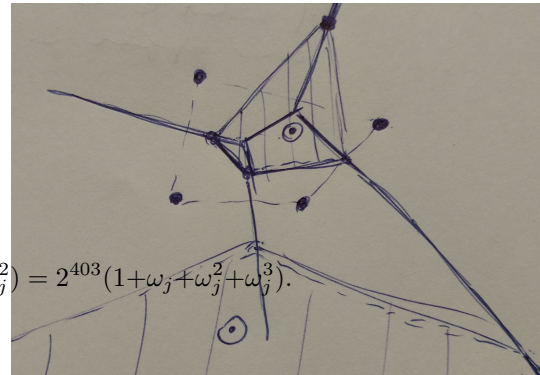
If the new point p is on the boundary, i.e. hitting one of the edges or vertices in $VD(X)$, then the same algorithm works. Just start with one of the incident regions.

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 leafs in G are created. Draw edges towards these leafs 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.
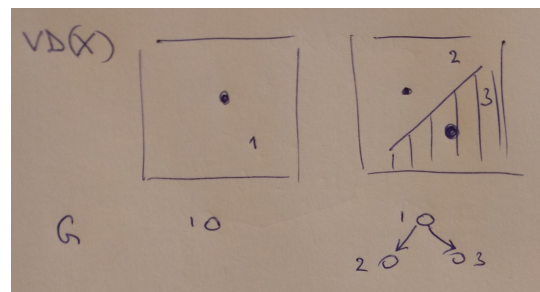
Expanding $f(x)$, we obtain:

$$f(x) = \frac{1}{2}\sum_{k=0}^{\infty}(2x)^k + \frac{1}{2}\sum_{k=0}^{\infty}(4x)^k,$$

and so, $a_k = \frac{1}{2}2^k + \frac{1}{2}4^k$. □

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 $b_1(x) = x^2 + x^3 + x^4 + \ldots = \frac{x^2}{1-x}$. The generating function for each other box is $b(x) = 1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x}$. 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

$$\binom{()}{n} - 2 + 4 - 1, 4 - 1) = \binom{()}{n} + 1, 3).$$

□

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

$$b_0(x) = \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = e^x - 1 - x$$

. 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

$$n!\left(\frac{4^n}{n!} - \frac{3^n}{n!} - n\frac{3^{n-1}}{n!}\right) = 4^n - 3^n - n3^{n-1}.$$

□
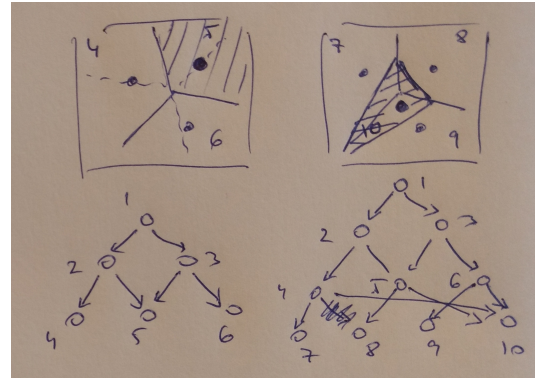
INCREMENTAL VORONOI DIAGRAM

Input: The set of points $X = \{v_1, \ldots, v_n\}$ 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. If the dual graph to $VD(X)$ (see 13.1.22 for the definition) is considered, then this is again a tessellation of the plane into convex regions. 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*.[13] Notice that it easy to turn any Delaunay tesselation into a triangulation by adding the necessary edges to triangulate the convex regions with

---

[13]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.

more edges. Any of these refined tesselations is called the Dalaynay 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 $CH(X)$. 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 τ and ν.

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*

$$\mathcal{A}(T) = (\gamma_1, \gamma_2, \ldots, \gamma_{3\tau}),$$

where γ_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.
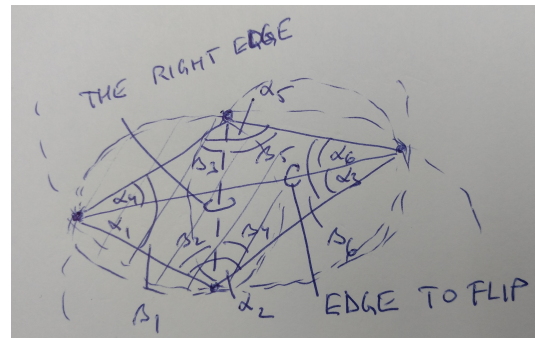
The flip is necessary if and only if one of the vertices outside the diagonal is inside the circle drawn through the remaining three vertices.

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:

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 $VD(X)$. Let $v_1, \ldots, v_k$ 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 trinagulation 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 $VD(X)$. □

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.
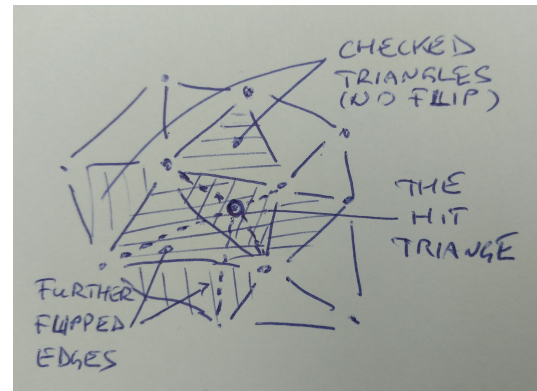
INCREMENTAL DELAUNAY TRIANGULATION

Input: The set of points $X = \{v_1, \ldots, v_n\}$ 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 Δ in $DT(X_k)$ hit by z (by the search in G)
  - if z is in the unbounded face, then
    - add the new triangles $\Delta_1, \ldots, \Delta_\ell$ to $DT(X_k)$ by joining z to visible edges in $CH(X_k)$
    - update the $CH(X)$.
  - if z hits a (bounded) triangle Δ, 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 $\Delta_1, \ldots, \Delta_4$ (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 Δ 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 kth 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 at al., cf. the link on page 971.

13.3.10. The beach line Voronoi. The Voronoi algorithm provides a perfect example for the the sweep line paradigm, where the extra structure to keep track of the events has to be quite smart.[14]

Imagine the horizontal line L (parallel to the x axis) flows in the top-down direction and meets the points in $X = \{v_1, \ldots, v_n\}$. 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,

$$R_L = \{p \in \mathbb{R}^2; \operatorname{dist}(p, L) \geq \operatorname{dist}(p, v_i), \ v_i \in X \text{ above } L\}.$$

This is exactly the part of $\mathbb{R}^2$ which can be tesselated 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.
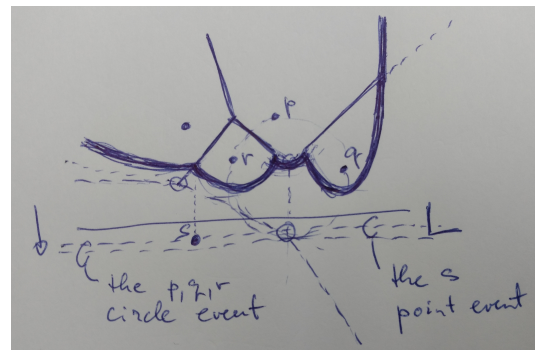
Call the boundary of R_L the *beach line* B_L. The vertices on B_L draw the $VD(X)$ when L is moving. Since the Voronoi diagram consists of lines, we do not even compute the parabolas and take care of the arrangements of the still active parts of parabolas in the beachline, as determined by the individual 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.

---

[14]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

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

$$\varphi(v) = \frac{1}{\|v\|^2} v = \frac{1}{x^2 + y^2}(x, y)$$
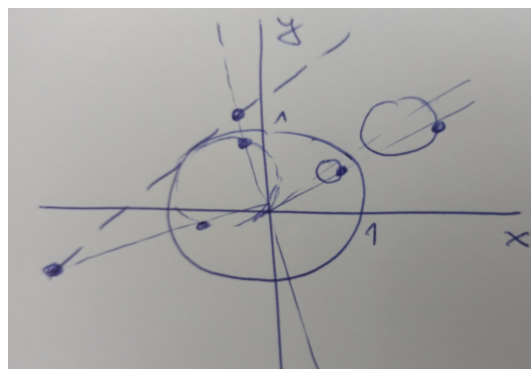
Clearly φ is a bijection on $\mathbb{R}^2 \setminus \{(0, 0)\}$. The geometric meaning of such a transform is clear from the formula, see the diagram. "The general" point v is sent to a point on the same line through the origin, but with reciprocal size. The unit sphere is the set of all fixed points.

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 φ.

Lemma. *The mapping φ 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.*

987

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)

$$\begin{aligned}
\|q - s\|^2 &= \|\frac{p}{\|p\|^2} - \frac{c}{\|c\|^2 - r^2}\|^2 \\
&= \frac{1}{\|p\|^2} + \frac{1}{(\|c\|^2 - r^2)^2}\|c\|^2 - 2\frac{p \cdot c}{(\|c\|^2 - r^2)\|p\|^2} \\
&= \frac{1}{(\|c\|^2 - r^2)^2}\|c\|^2 - \frac{1}{\|c\|^2 - r^2} \\
&= \left(\frac{r}{\|c\|^2 - r^2}\right)^2.
\end{aligned}$$

The latter computation assumes $\|c\| \neq r$. Fix the center c and consider diameter approaching r from below or above. Then the image is a circle with the fixed center s and a fast growing radius. In the limit position, the line is obtained as requested. (Check the computation directly yourself, if any doubts.) □

The continuity of φ has got important consequences. Consider a general plane μ (not containing the origin). The inversion φ maps one of the half-spaces determined by μ 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 $X = \{v_1, \ldots, v_n\}$ 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 end no four them lie on the same circle.

The spherical inversion φ maps the entire plane $z = 1$ to the sphere S with center $c = (0, 0, 1/2)$ and radius $1/2$. Write $w_1, \ldots, w_n$ for the images $w_i = \varphi(v_i)$.

Now, consider $CH(Y)$ for the set of the images $Y = \{w_1, \ldots, w_n\}$. 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 μ 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_μ with the plane $z = 1$. This is the defining condition for obtaining one of the vertices of the Voronoi diagram. Since the map φ preserves incidencies, 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). In the opposite direction, we demonstrate how handy continuous methods can be.

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

$(1, 1, 1, 1, 1, 0, 0, \dots)$ 1-crowns

$(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, \dots)$ 2-crowns

$(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, \dots)$ 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

$$a(x) = a_0 + a_1 x + a_2 x^2 + \cdots = \sum_{i=0}^{\infty} a_i x^i.$$

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 α 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 α^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 $(a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k)(b_0 + b_1 x + b_2 x^2 + \cdots b_k x^k)$. 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 the third part of chapter 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 $(1, 1, 1, \dots)$. From the sixth chapter, 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 i on page 340). 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:

$$\frac{1}{1-x} = \sum_{n \geq 0} x^n,$$

$$\ln(1+x) = \sum_{n \geq 1} (-1)^{n+1} \frac{x^n}{n},$$

$$\ln \frac{1}{1-x} = \sum_{n \geq 1} \frac{x^n}{n},$$

$$e^x = \sum_{n \geq 0} \frac{x^n}{n!},$$

$$\sin x = \sum_{n \geq 0} (-1)^n \frac{x^{2n+1}}{(2n+1)!},$$

$$\cos x = \sum_{n \geq 0} (-1)^n \frac{x^{2n}}{(2n)!}.$$

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 $a(x)$ and evaluate all the derivatives in $x = 0$. Obviously

$$a^{(k)}(0) = r(r-1)\cdots(r-k+1)(1+x)^r|_{x=0}$$
$$= r(r-1)\cdots(r-k+1)$$

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:

$$\frac{a_{k+1}\,x^{k+1}}{a_k\,x^k} = \frac{\frac{r(r-1)\ldots(r-n)}{(n+1)!}}{\frac{r(r-1)\ldots(r-n+1)}{n!}}\frac{x^{k+1}}{x^k} = \frac{r-n}{n+1}x.$$

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. Substituting $-x$ for the argument just kills the signs appearing in the generalized binomial coefficients. $\qquad\square$

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 $(a_0, a_1, a_2, \ldots)$, $\frac{1}{1-x}a(x)$ is the generating function for the sequence of all the partial sums $(a_0, a_0+a_1, a_0+a_1+a_2, \ldots)$. For instance,

$$\frac{1}{1-x}\ln\frac{1}{1-x}$$

is the generating function of the *harmonic numbers*

$$H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

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.

First, sort out the well-known example of the Fibonacci sequence, given by the recurrence

$$F_{n+2} = F_n + F_{n+1},\ F_0 = 0,\ F_1 = 1,$$

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 nth 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 $(0, F_0, F_1, F_2, \ldots)$, and $x^2F(x)$ does to $(0, 0, F_0, F_1, \ldots)$. Therefore, the generating function

$$G(x) = F(x) - x\,F(x) - x^2 F(x)$$

993

represents the sequence

$$(F_0, F_1 - F_0, 0, 0, \ldots, 0, \ldots).$$

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 a generating function for $1/(1 - x - x^2)$. Namely, write $r \in \mathbb{N}$

$$F(x) = \frac{1}{1 - x - x^2} = \frac{A}{x - x_1} + \frac{B}{x - x_2}$$
$$= \frac{a}{1 - \lambda_1 x} + \frac{b}{1 - \lambda_2 x},$$

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, λ_1, and λ_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 $\lambda_{1,2} = \frac{2}{1 \pm \sqrt{5}}$. 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 $F_1 = 2$.

<div align="center">METHOD TO RESOLVE RECURRENCES</div>

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 $a_{-1} = a_{-2} = \cdots = 0$).

(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.

$$\sum_{n \geq 0} a_n x^n = 5x \sum_{n \geq 0} a_{n-1} x^{n-1} - 6x^2 \sum_{n \geq 0} a_{n-2} x^{n-2}$$
$$- \sum_{n \geq 0} n x^n + 2x$$

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).

$$A(x) = 5xA(x) - 6x^2 A(x) - x \frac{1}{(1-x)^2} + 2x.$$

Step 3. Find 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

$$a_n = -\frac{1}{4} 3^n + 2^{n+1} - \frac{1}{2} n - \frac{7}{4}.$$

The first eight terms in the sequence are $0, 1, 3, 6, 8, -1, -59, -296$.

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.

$$B(x) = \sum_{n,k} b_k b_{n-k-1} x^n + \sum_{n,k} [n = 0] x^n$$

$$= \sum_{k} b_k x^k \left(\sum_{n} b_{n-k-1} x^{n-k} \right) + 1$$

$$= \sum_{k} b_k x^k (xB(x)) + 1 = B(x) \cdot xB(x) + 1.$$

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 $x \to 0_+$ is ∞, 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

$$(1 - 4x)^{1/2} = \sum_{k \geq 0} \binom{1/2}{k} (-4x)^k$$

$$= 1 + \sum_{k \geq 1} \frac{1}{2k} \binom{-1/2}{k-1} (-4x)^k.$$

Dividing $1 - \sqrt{1 - 4x}$ by the expression $2x$ leads to

$$B(x) = \sum_{k \geq 1} \frac{1}{k} \binom{-1/2}{k-1} (-4x)^{k-1}$$

$$= \sum_{n \geq 0} \binom{-1/2}{n} \frac{(-4x)^n}{n+1}$$

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 $L = (L[0], \ldots, L[n])$
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

$$L = \mathrm{Qsort}(L_1) + (L[0]) + \mathrm{Qsort}(L_2)$$

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 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:

$$(1) \qquad C_n = n - 1 + \sum_{k=1}^{n} \frac{1}{n} \left(C_{k-1} + C_{n-k} \right).$$

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

$$(2) \qquad nC_n = n(n-1) + 2 \sum_{k=1}^{n} C_{k-1}.$$

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

$$xC'(x) = \frac{2x^2}{(1-x)^3} + 2\frac{xC(x)}{1-x}.$$

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

$$\left((1-x)^2C(x)\right)' = \frac{2x}{1-x},$$

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

$$\sum_{k=1}^{n}\frac{1}{k}(n-k+1) = (n+1)\sum_{k=1}^{n}\frac{1}{k} - n$$
$$= (n+1)H_n - n = (n+1)(H_{n+1} - 1),$$

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 $(1,1,\ldots)$. 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:

$$e^x \overset{\text{e.g.f.}}{\longleftrightarrow} (1,1,1,\ldots),$$

$$\frac{1}{1-x} \overset{\text{e.g.f.}}{\longleftrightarrow} (1,1,2,6,24,\ldots,n!,\ldots)$$

$$\ln\frac{1}{1-x} \overset{\text{e.g.f.}}{\longleftrightarrow} (0,1,1,2,6,24,\ldots)$$

The slight modification of the definition (just forgetting about the $\frac{1}{n!}$ coefficient) is responsible for a very different behaviour, compared to the ordinary generating functions. 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.
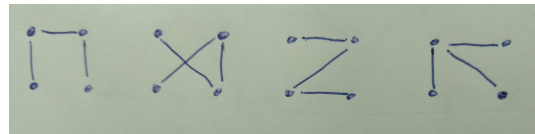
The recurrent formula can be obtained by fixing one of the vertices and add together the possibilities for all available degrees of this vertex. This suggests looking rather at the number R_n of the rooted trees. It is clear that $R_n = nT_n$ because there are n possibilities to place the root at each 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 $k_1 \ldots, k_m$ of the corresponding subtrees. Obviously $k_1 + \cdots + k_m = n - 1$, 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

$$R_n = n! \sum_{m>0} \frac{1}{m!} \sum_{k_1+\cdots+k_m=n-1} \frac{1}{k_1!\ldots k_m!} R_{k_1} \ldots R_{k_m}.$$

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.

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 $n > 0$,*

$$g_n = \lim_{w \to 0} \left(\frac{d^{n-1}}{dw^{n-1}} \left(\frac{w}{f(w)} \right)^n \right).$$

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

$$[x^n]\widehat{R}(x) = \frac{1}{n}[w^{n-1}] \left(\frac{w}{w/e^w} \right)^n$$

$$= \frac{1}{n}[w^{n-1}]e^{wn} = \frac{1}{n}\frac{n^{n-1}}{(n-1)!} = \frac{n^{n-1}}{n!}$$

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 $1, 2, \ldots, 6$ 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 $1, 2, \ldots, 7$ 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 $1, 2, \ldots, 5$ 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 $1, 2, \ldots, 5$ 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 $1, 2, \ldots, 6$ 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 $1, 2, \ldots, 6$ 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[4]. □

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. ◯

13.J.11. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:

◯
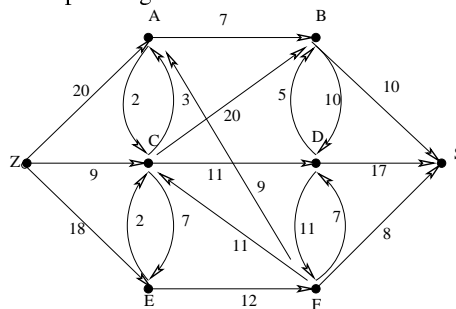
13.J.12. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:
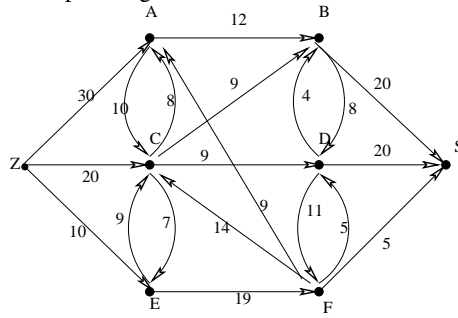
13.J.13. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:
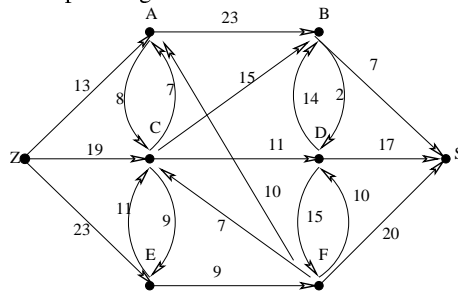
13.J.14. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:
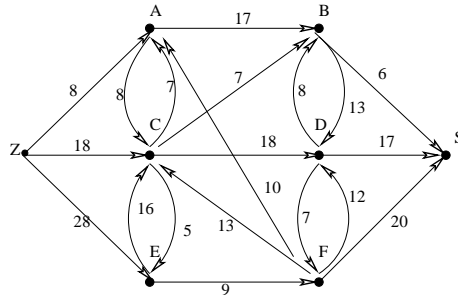
13.J.15. Find the generating functions of the following sequences:

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 $(1, 1, 1, 1, 1, \dots)$. 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. □

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? ○

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 $\binom{72}{2} - \binom{41}{2} - \binom{31}{2} - \binom{21}{2}$. □

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 $(x + x^2 + x^3 + x^4 + x^5 + x^6)^{12}$. ○

Solution. 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.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 $\binom{14}{3} - \binom{12}{3} - \binom{3}{3}$. □

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 $n = 1, 2, 3 \ldots$.

  ii) $a_0 = 0, a_1 = 1, a_{n+2} = 2a_{n+1} - 4a_n$ for $n = 0, 1, 2, 3 \ldots$.

Solution.

  i) $a_n = 2 + 3^{n-1}$.

  ii) $a_n = \frac{1}{2}\sqrt{-3} \cdot ((1 + \sqrt{-3})^n - (1 - \sqrt{-3})^n)$.

13.J.21. Solve the recurrence where each term of the sequence $(a_0, a_1, a_2, \ldots)$ is equal to the arithmetic mean of the preceding two terms.

Solution. $a_n = k\left(-\frac{1}{2}\right)^n + l$.

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 $b_n = \log_2 a_n$.

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).

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 $\frac{315}{5040} = \frac{1}{16}$.

$\square$

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. $(3, 1), (3, 2), (3, 4), (3, 5), (3, 6), (1, 2), (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), \ldots (5, 6)$.

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.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.11.
$$1 - \frac{4! \cdot 4!}{\frac{8!}{2^4}} = \frac{27}{35}$$

.

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 $(1, 2, 3, 4, \ldots)$ is $\frac{1}{(1-x)^2}$.
 ii) Since we have (by the previous exercise as well)

$$\frac{x}{(1 - x)^2} \overset{\text{o.g.f.}}{\longleftrightarrow} (0, 1, 2, 3, \ldots),$$

mĂĄme pro derivaci tĂŠto funkce

$$\left(\frac{x}{(1 - x)^2} \right)' = \frac{1 + x}{(1 - x)^3} \overset{\text{o.g.f.}}{\longleftrightarrow} (1 \cdot 1, 2 \cdot 2, 3 \cdot 3, \ldots).$$

Let us emphasize that this problem could also be solved using the fact that $\frac{1}{(1-x)^3} \overset{\text{o.g.f.}}{\longleftrightarrow} \binom{n+2}{n}$.

iii) We have

$$\frac{1}{1 - x} \overset{\text{o.g.f.}}{\longleftrightarrow} (1, 1, 1, 1 \ldots),$$

$$\frac{1}{1 - 2x} \overset{\text{o.g.f.}}{\longleftrightarrow} (1, 2, 4, 8, \ldots),$$

$$\frac{1}{1 - 2x^2} \overset{\text{o.g.f.}}{\longleftrightarrow} (1, 0, 2, 0, 4, 0, \ldots),$$

$$\frac{x}{1 - 2x^2} \overset{\text{o.g.f.}}{\longleftrightarrow} (0, 1, 0, 2, 0, 4, \ldots),$$

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

$$F(x) - x^2 F(x) + \tfrac{x}{1-x^3}.$$

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.

Index