MASARYK UNIVERZITY

Brisk guide to maths

Jan Slovák, Martin Panák, Michal Bulant
et al

Brno 2013

The work on the textbook has been supported by the project CZ.1.07/2.2.00/15.0203.

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitivness
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTMENTS IN EDUCATION DEVELOPMENT

Authors:
Mgr. Michal Bulant, Ph.D.
Mgr. Aleš Návrat, Dr. rer. nat.
Mgr. Martin Panák, Ph.D.
prof. RNDr. Jan Slovák, DrSc.
RNDr. Michal Veselý, Ph.D.

Graphics and illustrations:
Mgr. Petra Rychlá

Contents

Preface

This textbook follows the years of lecturing Mathematics at the Faculty of Informatics at Masaryk University in Brno. The programme requires introduction to genuine mathematical thinking and precision, but there is not much time dedicated.

Thus, we want to cover seriously, but quickly, about as much of mathematical methods as usual in bigger courses in the classical Science and Technology programmes. At the same time, we do not want to give up the completeness and correctness of the mathematical exposition. We want to introduce and explain more demanding parts of Mathematics, together with elementary explicit examples how to use the results. But we do not want to solve for the reader how much of theory or practice to enjoy and in which order.

All these requests have lead us to the two collumn format where the rahter theoretical explanation and the practical examples are split. This way, we want to please 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 everything in a linear order. On the opposit, the readers should enjoy browsing through the text and finding their own paths.

In both collumns, we intend to present rather standard exposition of basic Mathematics, but focusing on the essence of the concepts and their relations. The examples are solving simple mathematical problems but we also try to show thei use in mathematical models in practise as much as possible.

We are aware thath the theoretical text is written in a very compact way. A lot of details are left to readers, in particular in the more difficult paragraphs. Similarly, the examples display the variety from very simple ones to those reuqesting some deeper thoughts.

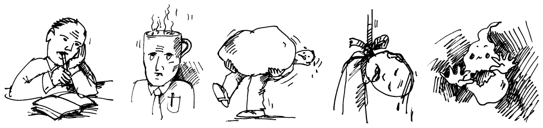
We would very much like to help the reader

- 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 creating mathematical models and to appretiate their usage.

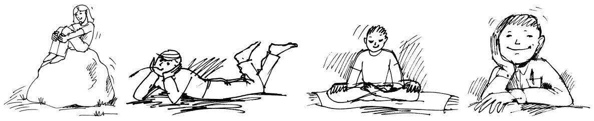
The goals are ambitions and nearly everyone needs his or her own paths, including failures. This is one of the reasons why we come back to basic ideas and concepts several times with growing complexity and width of the discussions. Of course, this might also look as chaotic, but we very much hope that this approach gives a much better chance to those who will persit in their effort.

Clearly this textbook cannot be the only source for everybody. Actually, the only really good proceeding is the combine several sources and to think about their differences on the way. But we hope, it should be a perfect begin and help for everybody who is ready to return back to the individual parts again and again.

To make this task simpler, we have added emotive icons. We hope they will not only spirit the dry mathematical text but also indicate which parts should be rather be read carefully or better jumped over 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. Roughly speaking, we are using icons warning before complexity, difficulty etc.:



Further icons indicated unpleasant technicality and need of patiance:



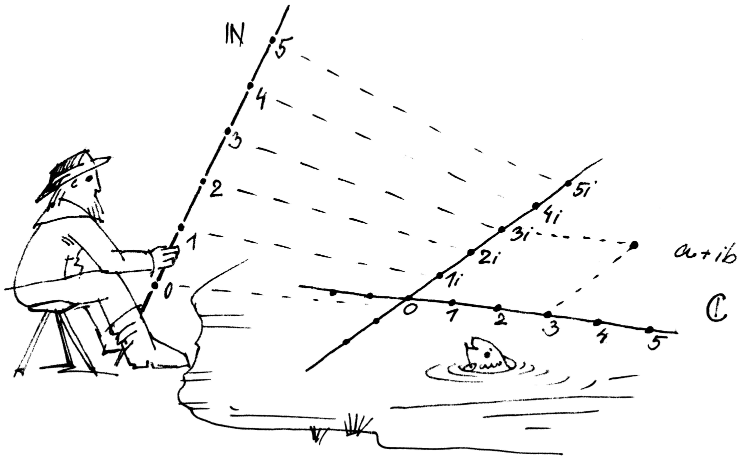
Finally, there also icons showing up the joy of the game:

The practical collumn with the examples should be readable nearly independently off the theory. Without the ambition to know the deeper reasons why the algorithms work, it should be possible to read just here. Some definitions and descriptions in the theoretical text are marked to be catched easily when reading the examples, too. The examples and theory are partly coordinated to allow jumping there and back, but the links are not tight.

CHAPTER 1

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 argue why rational numbers are not sufficient for us (although computers are actually not able to work with any other) and we recall the so-called complex numbers (because even reals are not enough for some calculations).

1.1. Find some real number which is not rational.

Solution. One among many possible answers is $\sqrt{2}$. Already the old Greeks knew that if we prescribe the area of rectangle $a^2 = 2$, then we cannot find rational a which would satisfy it. Why?

Assume we know that it holds $(p/q)^2 = 2$ for natural numbers p and q that do not have common divisors different from 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, and so is p. Hence, p^2 is divisible by 4, and so q must be even that implies p and q have 2 as a common factor, which is a contradiction. □

1.2. Remark. It can be even proven that n-th root of a rational number, where n is natural, is either natural or is not rational (see ||G||

The goal of the first chapter is to introduce the reader to the fascinating world of mathematical thinking. For that we choose our examples of mathematical modelling of real situations using abstract object and connections to be as specific as possible. We also go through a few topics and mechanisms to which we will subsequently return in the rest of the book, and in the end of the chapter we will speak about the language of mathematics itself (which we will mostly use in an intuitive way).

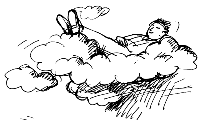
The easier the objects and settings we work with are, the more it is difficult to understand in depth the nuances of the use of particular tools and mechanisms. Mostly it is possible to reach the core ideas only through their connection to others. Therefore we introduce them from many points of view at once.

Changing the topics very often might be confusing, but it will surely get better when we return to specific ideas and notions in later chapters.

The name of this chapter can be also understood as an encouragement to patience. Even the simplest tasks and ideas are easy only for those who have already seen similar ones. Full knowledge and mathematical thinking can be reached only through a long and complicated.

Let us start with the simplest thing: common numbers.

1. Numbers and functions



Since the dawn of ages people wanted to know "how much" of something they had, or "how much" is something worth, "how long" will a particular task take, etc. The result of such ideas is usually some "number". We consider something to be a number, if we can multiply it and add it, and it behaves according to the usual rules – either according to all rules we except, or only to some. For instance, the result of multiplication does not depend on the order of multiplicands, we have the number zero whose adding does not change the result, we have the number one which behaves in a similar manner with respect to addition, and so on.

The simplest example are the so-called natural numbers, which we denote $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. Note that we consider zero to be a natural number, as it is usual especially in computer science.

To count "one, two, three, …" is learned already by little children in their pre-school age. Some time later we meet the integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ and finally we get used to floating-point numbers, and we know what a 1.19-multiple of the price means thanks to the 19% tax.

1.3. Find all solutions to the equation $x^2 = b$ for any real number b.

Solution. 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 real x cannot exist. Thus we need to find a bigger domain, where this equation has a solution.

First we add to the real numbers a new number i, so-called *imaginary unit*, and try to extend the definitions of addition and multiplication in order to preserve the usual behaviour of numbers (as summarised in 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 $z = a + i\,b$.

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 tuples of real numbers, with the additional rule that $i^2 = -1$, that is,

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

□

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 $\operatorname{re}(z) = a$, $\operatorname{im}(z) = b$.

1.4. Assert that all the properties (KG1-4), (O1-4) and (P) of scalars from 1.1 hold.

Solution. 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. All properties are obtained by direct calculations. □

Complex number is given by a tuple of real numbers, therefore it is a point in the real plane $\mathbb{R}^2$.

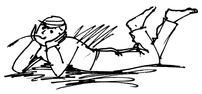
1.5. Show that the distance of the complex number $z = a + i\,b$ from the origin (we denote it by $|z|$) is given by the expression $z\bar{z}$, where $\bar{z}$, the *complex conjugate*, is $a - i\,b$.

Solution. The product

$$z\bar{z} = (a^2 + b^2) + i\,(-ab + ba) = a^2 + b^2$$

is always a real number and indeed gives us the square of the distance from the number z to the origin. Thus it holds $|z|^2 = z\bar{z}$. □

1.1. Properties of numbers. In order to be able to properly work with numbers, we need to be more 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*. A good choice of axioms determines the range of the theory they give rise to, and also to its usability in mathematical models of reality.



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

Properties of scalars

Properties of numbers:

(KG1) $(a + b) + c = a + (b + c)$, for all a, b, c

(KG2) $a + b = b + a$, for all a, b

(KG3) there exists 0 such that for all a it holds that $a + 0 = a$

(KG4) for all a there exists b such that $a + b = 0$

The properties (KG1)-(KG4) are called the properties of *commutative group*. They are called *associativity*, *commutativity*, *existence of neutral element* (when speaking of addition we usually say zero element), *existence of inverse element* (when speaking of addition we also say the negative of a and denote it by $-a$), respectively.

Properties of multiplication:

(O1) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, for all a, b, c

(O2) $a \cdot b = b \cdot a$, for all a, b

(O3) there exists 1 such that for all a it holds that $1 \cdot a = a$

(O4) $a \cdot (b + c) = a \cdot b + a \cdot c$, for all a, b, c.

The properties (O1)-(O4) are called *associativity*, *commutativity*, *existence of unit element* and *distributivity* of addition with respect to multiplication, respectively.

The sets with operation $+$, $\cdot$ that satisfy the properties (KG1)-(KG2) and (O1)-(O4) are called *commutative rings*. Further properties of multiplication:

(P) for every $a \neq 0$ there exists b such that $a \cdot b = 1$.

(OI) if $a \cdot b = 0$, then either $a = 0$ or $b = 0$

The property (P) is called *existence of inverse element* with respect to multiplication (this element is then denoted by a^{-1}) and the property (OI) then says that there exists no "divisors of zero".

Properties of the operations of addition and multiplication will be often used, even if we do not know what object are we really working with. In this way we obtain very general mathematical tools. However, it is good to have some idea of typical examples of object we work with.

The integers $\mathbb{Z}$ are a good example of commutative group, the natural numbers are not since they do not satisfy (KG4) (and possibly do not even contain the neutral element if one does not consider zero to be natural).

If a commutative ring also satisfies the property (P), we speak of *field* (often also about *commutative field*).

1.6. Remark. The distance $|z|$ is also called the absolute value of the complex number z.

1.7. Polar form of complex numbers. Let us first 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 z to the origin (measured in the positive sense and taking values between 0 and 2π to avoid ambiguity). These numbers describe exactly all points on the unit circle in the complex plane. Every non-zero number z can be then written in a unique way as

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

The number φ is called the argument of complex number z.

1.8. Multiplication of complex numbers in the polar form. Let the numbers $z_1 = |z_1|\,(\cos(\varphi_1) + i\ \sin(\varphi_1))$ and $z_2 = |z_2|\,(\cos(\varphi_2) + i\ \sin(\varphi_2))$ be given, and let us calculate their product:

$$\begin{aligned} z_1 \cdot z_2 &= |z_1|\,(\cos(\varphi_1) + i\ \sin(\varphi_1)) \cdot |z_2|\,(\cos(\varphi_2) + i\ \sin(\varphi_2)) \\ &= |z_1||z_2|\Big[\cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2) + \\ &\quad +i\ (\cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2))\Big] \\ &= |z_1||z_2|\,[\cos(\varphi_1 + \varphi_2) + i\ \sin(\varphi_1 + \varphi_2)]\,, \end{aligned}$$

the last equality is a result of addition formulas for trigonometric functions. Repeated application of the previous formula on the product of the number z with itself yields the so-called "Moivre theorem":

$$z^n = [|z|(\cos\varphi + i\sin\varphi)]^n = |z|^n(\cos(n\varphi) + i\sin(n\varphi)).$$

1.9. Express the number $z_1 = 2 + 3i$ in the polar form and express the number $z_2 = 3(\cos(\pi/3) + i\sin(\pi/3))$ in the algebraic form.

Solution. The absolute value of the given number (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 picture we can easily compute $\sin(\varphi) = 3/\sqrt{13}$, $\cos(\varphi) = 2/\sqrt{13}$. Thus it is $\varphi = \arcsin(3/\sqrt{13}) = \arccos(2/\sqrt{13}) \doteq 53,3°$. In total,

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

Transition from the polar form to algebraic is even simpler:

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

□

The last stated property (OI) is automatically satisfied if (P) holds. However, it does not work the other way and thus we say that the property (OI) is weaker than (P). For instance the ring of integers $\mathbb{Z}$ does not satisfy (P) but satisfies (OI). In such case we use the term *integral domain*.

Let us note that the set of all non-zero elements in the field along with the operation of multiplication satisfies (O1), (O2), (O3), (P) and thus is also a commutative group. Only, instead of addition we speak of multiplication. As an example, we can take all non-zero real numbers.

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

All properties (KG1)-(KG4), (O1)-(O4), (P), (OI) from our consideration are necessary to be considered an *axiomatic definition* of the corresponding mathematical terms. For our needs it is enough to keep in mind that in further discussions we will use only these properties of scalars and thus our results will hold for any objects with such properties.

Exactly this is the true power of mathematical theories – they do not hold just for a specific solved example. Quite the opposite, when build in a rational way they are always universal. We will try to emphasise this aspect, although our ambitions are very humble due to the limited size of the book.

1.2. Existence of scalars. In order to build a mathematical theory, we need to ensure that such objects can exist. Thus we will show how to construct basic sets of numbers. For the construction of natural numbers let us start with the assumptions that we know what sets are.

Let us denote the empty set by $\emptyset$ and define

$$0 := \emptyset, \quad n + 1 := n \cup \{n\}\ , \tag{1.1}$$

in other words

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

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

In this way we identify the natural numbers with the number of elements of some specific sets. The number n is a set with exactly n elements, and two natural numbers a, b are identical if and only if the corresponding sets have the same number of elements. In the set theory we say "cardinality" of a set instead of "the number of elements" of a set. This term makes sense even for infinite sets (the other does not).

At first sight we also see the usual definition of ordering of natural numbers according to their size (the number a is strictly smaller then b if and only if $a \neq b$ and $a \subset b$ as a set). A further formal step should be the definition of addition and multiplication and a proof of all basic properties of natural numbers, including the afore-stated axioms of commutative ring. For instance, we can easily show that every subset in $\mathbb{N}$ has a smallest element and many other properties about which we usually don't think too long and consider them trivial.

We will not deal with the construction of numbers deal in much detail and assume that the reader knows the rational numbers

1.10. 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 out its absolute value and argument. Let us first calculate the absolute value:

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

Now for the argument φ we have:

$$\cos\varphi = \tfrac{\mathrm{re}(z)}{|z|} = \tfrac{1+\frac{1}{2}}{\sqrt{3}} = \tfrac{\sqrt{3}}{2}, \quad \sin\varphi = \tfrac{\mathrm{im}(z)}{|z|} = \tfrac{1}{2},$$

therefore $\varphi = \pi/6$. Thus we have obtained

$$z = \sqrt{3}\left(\cos \tfrac{\pi}{6} + i \sin \tfrac{\pi}{6}\right).$$

□

1.11. Using the Moivre theorem calculate

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

Solution. We immediately obtain

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

□

More examples about complex numbers can be found at ||1.111||

1.12. Complex numbers are not just a tool to obtain "weird" solutions to quadratic equations, but 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. Let us 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 let us introduce unknowns u, v satisfying conditions $u + v = t$ and $3uv + p = 0$. Plugging the first condition to the previous equation we obtain

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

and then plugging the second yields

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

which is a quadratic equation in the unknown $s = u^3$. Thus we have

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

($\mathbb{Q}$), the real numbers ($\mathbb{R}$) and the complex numbers ($\mathbb{C}$). From time to time we will recall some theoretical and practical connections during the course of the book. In detail, the construction of rational numbers from the natural numbers is done in 1.40. The construction of real numbers is closely connected to the limit processes which we do later, and some time sooner we will deal with complex numbers from various algebraic viewpoints. The figure depicts how can we perceive number domains as embedded one into another (that is, the complex plane contains many copies of natural and integer numbers, the real line, and so on).

Furthermore, as it 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.

1.3. Scalar functions. We often work with some value which is not given as a particular number. Instead, we something about the dependence of the value on other values. Formally we write that the value $y = f(x)$ of our "dependent" variable value y is give by the "independent" value x. We can consider the knowledge of f formally (it is just some unspecified dependence) or operationally – $f(x)$ is given by a formula which consists of known operations (for now, we assume a there is just a finite number). If the value is scalar, we speak of a *scalar function.* Every function is defined over some set, we speak of *domain* of a function, and the range of all values of the function is then the so-called *codomain* of the function.

We can have the values of f to be given only approximately or with some probability.

The point of mathematical considerations is then to derive, from a non-formal description of dependencies, explicit formulas for functions which describe them, or at least explicit evaluations for specific values of dependent values, or their approximation. According to the type of he task and the goal we work with

- exact finite expression,
- infinite expression,
- approximation of an unknown function with a known value (usually with explicitly evaluated possible error of the approximation),
- approximation of values with evaluation of their probability, etc.

Scalar function is for instance yearly pay of a worker in some company (the values of of independent variable, that is, the domain of the function, are individual workers x from the set of all considered workers, $f(x)$ is their yearly pay for the given year). Analogously, we can observe monthly pay of a particular worker in time (the independent variable is then time in months, dependent variable is the pay in the given month). Another example is then the area of a planar object, volume of an object in space, speed of a particular car in time, etc. We can surely imagine that in all given cases the value can be given by some loosely described connection or measured approximately, etc.

1.4. Functions defined by operations. Functions can be given by listing their values – for instance in a company there are only finitely many employes and we can compose a table with their actual monthly pays. More often, we have some rules how to obtain the values instead of definite tables.

By substituting back, we obtain

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

In the expression for u there is cubic root, and 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 (‖**??**‖)). All these three solutions are called cubic root of a. Therefore the expression $\sqrt[3]{a}$ has three meanings in the complex domain. If we want to have a single meaning for that expression, we usually consider it to be the solution with the smallest argument.

Let us further add that in the used method there could possible arise division by zero. In this case another method (usually simpler) is necessary.

1.13. Solve the equation

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

Solution. As we can easily find out, this equation has no rational roots (methods to determine rational roots will be introduce in the part (‖**??**‖)). Plugging into obtained formulas we yield $p = b - a^2/3 = -7/3$, $q = -7/27$, for u we then have

$$u = \frac{\sqrt[3]{28 \pm 12\sqrt{-147}}}{6},$$

where 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 as we can easily see, we obtain only three distinct values for x. Plugging into (‖1.12‖) 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 we have said before, we see that even if we have used complex number during the course, the result is real. □

B. Combinatorics

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

Factorial function

An important scalar function defined over natural numbers is the so-called *factorial*, defined by the relation

$$f(0) = 1, \ f(n) = n \cdot f(n-1)$$

for $n = 1, 2, \dots$. We write $f(n) = n!$ and the definition clearly means that

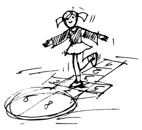
$$n! = n \cdot (n-1) \cdots 1.$$

Our definition of the factorial function says how its value $f(n)$ changes when we change the value of n by one. The formula for $n!$ then explicitly says, how much is that in total. In this case it is not a very effective formula, since its complexity increases with increasing n, but it is hard to find a better one.

Let us have a look on the natural number addition as on a scalar function defined by operations. The domain is then the set of all tuples (a, b) of natural numbers. We define $a + b$ as a result of the procedure of adding 1 to a a couple of times. Recall that we have defined $a + 1$ in general in the equation (1.1). For every addition of one we remove the greatest element from b, and we carry this on until b is not empty (effectively, for every addition we subtract one from b, whose value at any given time tells us how many more ones shall we add).

It is evident that addition defined in this way is given by an (iterative) formula, but this approach is not very efficient for practical addition. This will happen in our exposition often – theoretical correct definition of a term or an operation does not mean that the given procedure can be carried effectively. For this reason we will develop theories to obtain practical tools. As for natural numbers, we are able to deal with them quickly (while they are small), and for the large ones we know the algorithm for reducing it to addition of more smaller numbers, and for the really large ones we have computers (unless they are very very large).

2. Combinatorics



A typical "combinatorial" problem is to enumerate in how many ways something can happen. For instance, in how many ways can we choose two different sandwiches from the daily offer in a grocery shop? Is this the situation where we consider that all sandwiches that are there are pairwise distinct or do we consider only distinct kinds of sandwiches? Do we then allow to take two of the same type? Many such question occur in the context of card and other games.

When solving particular problems we usually use either the so-called "rule of product", where in mutually independent tasks we combine all possible results, or the "rule of sum", where we sum the number of ways for distinct incompatible tasks. Practically we demonstrate it in many examples.

1.5. Permutations. If from a set of n elements we create some order of the elements, we can choose the first element in n ways, the second can be then chosen in $n - 1$ ways and so on, until we take the last element for which there is only one choice. Therefore for a given finite set S with n elements there are exactly $n!$ distinct orders (we have used the product rule). The process of ordering elements of a set S is called *permutation* of the elements of S. The result of a permutation is always some ordering of the elements.

1.14. 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 both pears and apples to be indistinguishable. We also allow that one of the children might not get anything.)

Solution. Five pears can be divided in six ways (it is determined by the number of pears given to John, the rest goes to Mary.) Six apples can be independently divided in seven ways. Using the rule of product, the total number is $6 \cdot 7 = 42$. □

1.15. Determine the number of four-digit numbers, which 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.

Solution. The set of the numbers described in the statement consists of two disjoint sets, that is, of numbers which start with the digit 1 but do not end with the digit 2 (the first set), and of numbers that do not begin with the digit 1 and end with the digit 2 (the second set). The total number is then obtained using the rule of sum 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 third digit in nine ways. These three choices then uniquely determine a number and using the rule of product we then have $10 \cdot 10 \cdot 9 = 900$ 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). Using the rule of sum we have $900 + 800 = 1700$ numbers. □

1.16. Determine the number of ways of placing white tower and black tower on the chessboards (of size 8×8), that they don't threat each other (that is, they are neither in the same column nor in the row).

Solution. Let us first place the white tower. For it we can choose among 8^2 positions. In the second step we place the black tower. Now we have "to our disposal" 7^2 positions. Using the rule of some the total number of ways is $8^2 \cdot 7^2 = 3\,136$. □

In the following examples we will use the notions of combination, permutation, variation (possibly with repetitions), which we have defined.

1.17. During the 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. Let us denote the two given speakers as A and B. If right after the speaker A follows B, we can see it as a speech of a single

If we first number the elements in S, that is, we identify S with the set $S = \{1, \dots, n\}$ of n natural numbers, then permutations correspond to possible orderings of 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:*

$$p(n) = n! \tag{1.2}$$

1.6. Combinations and variations. Another simple example of a value determined by formula are so-called *binomial coefficients*, which express in how many ways can we choose k distinguishable items from a set of n items. Clearly we have

$$n(n-1)\cdots(n-k+1)$$

of possible results of a subsequential choosing of our k elements, but we obtain the same k-tuple in $k!$ distinct orders. If we want to choose the items along with an ordering, we speak of a *variation* of k-th degree.

As we have just checked, the number of combinations and variations are given by the following formula, which are not very effective for calculations with k and n large, since they contain factorials.

Combinations and variations

Proposition. *For the number $c(n,k)$ of combinations of k-th degree among n elements, where $0 \le k \le n$, it holds that*

(1.3)
$$c(n,k) = \binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots 1} = \frac{n!}{(n-k)!k!}.$$

For the number $v(n,k)$ of variations it holds that

$$v(n,k) = n(n-1)\cdots(n-k+1) \tag{1.4}$$

for all $0 \le k \le n$ (and zero otherwise).

We pronounce binomial coefficient $\binom{n}{k}$ as "n over k". The name stems from the so-called *binomial expansion*, which is the expansion of $(a+b)^n$. If we expand $(a+b)^n$, the coefficient at $a^k b^{n-k}$ equals for every $0 \le k \le n$ exactly the number of ways to choose a k-tuple from n parentheses in the product (from these parentheses, we take a, from the others, we take b). Therefore we have

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \tag{1.5}$$

and note that for the derivation only distributivity, commutativity and associativity of multiplication and summation was necessary. The formula (1.5) therefore holds in every commutative ring.

Let us present another simple example of a mathematical proof – a few simple propositions about binomial coefficients. For a simplification of formulations we define $\binom{n}{k} = 0$ whenever $k < 0$ or $k > n$.

1.7. Proposition. *For all natural numbers k and n we have* speaker AB. The number of all orderings where B speaks right after A is then equal to the number of permutations of seven elements. Clearly, the same number is for the number of all orderings where A speaks right after B. Since the number of all possible orderings of eight speakers is 8!, the result is $8! - 2 \cdot 7!$. □

1.18. How many anagrams 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 distinct letters and there are 6! words of six indivisible letters. In our case we have to multiply this by two, since are double-letter can be either BR or RB. Thus the result is $2 \cdot 6!$.

b) $7! - 2 \cdot 6!$ the complement to the part a) to the number of all seven-letter words of distinct letters. □

1.19. In how many ways can an athlete place 10 distinct cups to 5 shelves, if into every shelf all 10 cups fit?

Solution. Let us add 4 indistinguishable items, say separators, to the cups. The number of all distinct orderings of cups and separators is clearly $14!/4!$ (the separators are indistinguishable). Every 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 number $14!/4!$ is the result. □

1.20. Determine the number of four-digit numbers with exactly two distinct digits.

Solution. Two distinct letters used for the number can be chosen in $\binom{10}{2}$ ways, from two chosen digits we can compose $2^4 - 2$ distinct four-digit numbers (we subtract the 2 for the two one digit numbers). 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 we have $630 - 63 = 567$ numbers. □

1.21. Determine the number of even four-digit numbers composed of exactly two distinct digits.

Solution. Analogously to the previous example, let us first ignore the peculiarities of the digit zero. We thus obtain $\binom{5}{2}(2^4-2)+5\cdot5(2^3-1)$ numbers (we first count only the number that consist only of even digits, the second summand gives the number of even four-digit numbers

(1) $\binom{n}{k} = \binom{n}{n-k}$
(2) $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$
(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 proposition is immediate directly from the formula (1.3). If we expand the right-hand side of (2), we obtain

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

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

In order to prove (3), we use the so-called *mathematical induction*. This tool is very suitable for statements saying that something should hold for every natural number n. The mathematical induction consists of two steps. In the first, base step we assert the claim for $n = 0$ (in general, for the smallest n the claim should hold for). In the second, inductive step we assume that the claim holds for some n (and all smaller numbers) and using this we prove that that this implies the claim for $n + 1$. Putting it together, we obtain that the claim holds for every n.

The claim (3) clearly holds for $n = 0$, since $\binom{0}{0} = 1 = 2^0$. (Similarly easy is it also for $n = 1$.) Now let us assume that the claim holds for some n and calculate the corresponding sum for $n + 1$ using the claims (2) and (3). We yield

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

Note that the formula (3) gives the number of all subsets of an n-element set, since $\binom{n}{k}$ is the number of all subsets of size k. Note also that (3) follows from (1.5) by choosing $a = b = 1$.

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

$$\begin{aligned}\sum_{k=0}^{n+1} k\binom{n+1}{k} &= \sum_{k=0}^{n+1} k\left[\binom{n}{k-1} + \binom{n}{k}\right]\\ &= \sum_{k=-1}^{n} (k+1)\binom{n}{k} + \sum_{k=0}^{n+1} k\binom{n}{k}\\ &= \sum_{k=0}^{n}\binom{n}{k} + \sum_{k=0}^{n} k\binom{n}{k} + \sum_{k=0}^{n} k\binom{n}{k}\\ &= 2^n + n2^{n-1} + n2^{n-1} = (n+1)2^n.\end{aligned}$$

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

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

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

□

1.22. There are 677 people at a concert. Do some of them have the same 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. □

1.23. New players meet in a volleyball team (6 people). How many times do they shake hands when introducing to each other (everybody shakes with everybody)? How many times do they shake hands with the opponent after playing a match?

Solution. Every tuple of players shakes hands at the introduction. The number of handshakes is then equal to the combination $C(2,6)=\binom{6}{2}=15$. After a match each of the six players shakes hands six times (with each of six opponents). Thus the number is $6^2=36$. □

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

Solution. On 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 driver plus the passengers can be seated among the 24 seats arbitrarily. Let us first choose the seats to be occupied, that is, $\binom{24}{21}$, and among these the people can be seated in 21! ways. That makes $2.\binom{24}{21}21! = \frac{24!}{3}$ ways. □

1.25. 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. Let us first compute the number of insertions ignoring the non-emptiness condition. Using the rule of product (we insert the 10-bills and 100-bills independently) we have $C(2,7)^2=\binom{7}{2}^2$. Let us now subtract the insertions such that exactly one envelope is empty and then the insertions such that two are empty. We have $C(2,7)^2-3(C(1,6)^2-2)-3=\binom{7}{2}^2-3(6^2-2)-3=336$. □

The second property from our claim allows us to compose all binomial coefficients into the so-called *Pascal triangle*, where every number is obtained as a sum of two coefficients situated right "above" it:

$$\begin{array}{lccccccccccc}
n=0: & & & & & & 1 & & & & & \\
n=1: & & & & & 1 & & 1 & & & & \\
n=2: & & & & 1 & & 2 & & 1 & & & \\
n=3: & & & 1 & & 3 & & 3 & & 1 & & \\
n=4: & & 1 & & 4 & & 6 & & 4 & & 1 & \\
n=5: & 1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Note that in individual rows we have exactly the coefficients at individual powers in the expression (1.5), for instance the last given row says

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

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

Let there be among n given elements p_1 elements of first kind, p_2 elements of 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 $P(p_1,\dots,p_k)$.

Similarly to permutations and combinations without repetitions, for the choice of the first element we have n possibilities, for the second $n-1$ and so on, until the last element, for which we have only one choice. But we consider the orderings which differ only in the order of indistinguishable elements to be identical. Elements of every kind can be ordered in $p_i!$ ways, thus we have

Permutations with repetitions

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

Free choice of k elements from n elements, when order matters, is called *variation of k-th degree with repetitions*, the number of those is denoted $V(n,k)$. Free choice in this case means that we assume that for every choice we have the same number of possibilities – for instance, when we return the elements back before the next choice, when we throw the same dice, and so on. The following clearly holds:

Variations with repetitions

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

If we are interested in choice without taking care of order, we speak of *combinations with repetitions* and for their number we write $C(n,k)$. At first sight, it does not seem to be easy to determine the number. The proof of the following theorem is typical for mathematics – we reduce the problem to another problem we have already dealt with. In our case it is reduction to standard combinations without repetitions:

Combinations with repetitions

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

$$C(n,k)=\binom{n+k-1}{k}.$$

1.26. Determine the number of distinct sentences which can arise by permuting letters in the individual words in the sentence "Skokan na koks" (the arising sentences and words do not have to make any sense).

Solution. Let us first compute the number of anagrams of individual words. From the word "skokan" we obtain $6!/2$ distinct anagrams (permutation with repetition $P(1, 1, 1, 1, 2)$), similarly "na" yields two and "koks" $4!/2$. Therefore, using the rule of product, we have $6!4!/4 = 4320$. □

1.27. How many distinct anagrams 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

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

□

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

Solution. Let us 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 three colours for every hole), that is variations with repetitions (see). We have

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

Now let us 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. □

1.29. In how many ways could have the English Premier League league finished, if we know that no two of the three teams Newcastle United, Fulham and Tottenham Hotspur are not "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 we subtract the tables

Proof. The proof is based on a trick (a simple one, as soon as we understand it). We show two different approaches.

Assume first, that we are drawing cards from a deck of n different cards and in order to make it possible to draw some card multiple times, we add to the deck $k - 1$ different jokers (we definitely want to draw at least one of the original cards). Say that we have drawn r original cards and s jokers, that is, $r + s = k$. It seems that we should devise a method how to assign "substitute" jokers to original cards, so that we know how many times we have drawn each original card. But we actually need to discuss only the number of ways how to do that.

For that we can use the mathematical induction and assume that the claim holds for any arguments smaller than n and k. We need to obtain the combination with repetition of the s-th order from r original cards, which gives $\binom{r+k-r-1}{s} = \binom{k-1}{s}$, which is exactly the number of combinations without repetitions of s-th order from all jokers. Thus the theorem is proven.

Alternative approach (induction-free): Over the set

$$S = \{a_1, \dots, a_n\},$$

from which we choose the combination, we fix an ordering of the elements and for our choices of elements of S we prepare n boxes into which we give (in the fixed order) the elements of S (one element into every box).

The individual choices $x_i \in S$ are then given to the box which already contains this element. Now let us realize that in order to detect the original combination we just need to know how many elements are there in individual boxes. For instance,

$$a \mid bbb \mid cc \mid d \simeq * \mid * * * \mid ** \mid *,$$

determines the choice b, b, c from the set $S = \{a, b, c, d\}$.

In the general case of the choice of k elements from n possible we have a chain of $n + k$ elements and the number $C(n, k)$ equals the number of possible placements of the boxes | among individual elements. This amount to the choice of $n-1$ positions from $n+k-1$ possible. Since we have

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

the theorem is proven (for the second time). □

3. Difference equations

In the previous paragraphs we have seen formulas, which determined the value of a scalar function defined on natural numbers (factorial) or on tuples of natural numbers (binomial coefficients) using already defined values. In the paragraph 1.5 the binomial coefficients are defined with a directly computable formula, but we can also understand them using the relationship exhibited in 1.8 – instead of the value of the function we give the difference corresponding to a change of the variable.

Such approach can be seen very often when formulating mathematical models that describe real systems in economy, biology, etc. We will observe only a few simple examples and we will return to this topic in the future.

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

$$20! - \binom{3}{2} \cdot 2! \cdot 19! + 3! \cdot 18! = 1741445647958016000.$$

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

$$\binom{18}{3} \cdot 17! \cdot 3! = 1741445647958016000.$$

ways. □

1.30. For any fix $n \in \mathbb{N}$ determine the number of all solutions to the equation

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

on the set of strictly positive integers.

Solution. Every solution $(r_1, \dots, r_k)$, $\sum_{i=1}^{k} r_i = n$ 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, $\binom{n+k-1}{n}$. □

C. Difference equations

Difference equations (also called recurrence relations) are relations between elements of some sequence, where an element of the sequence depends on previous elements. To solve a difference equations means finding an explicit formula for n-th (that is, arbitrary) element of the sequence. Recurrence relation allows us only to compute n-th element by computing all previous elements.

If an element of the sequence is determined only by the previous element, we speak about first order difference equation. Those are present in our real world, 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 find out how much shall we pay per month if we want to repay a loan in a fixed time.

1.31. 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 to buy the car with yearly interest of 6%. Michael would like to finish repaying the loan in three years. How much should he pay per month?

1.9. Linear difference equations of first order. General *difference equation of the first order* is an expression of the form

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

where F is a known scalar function with two parameters. 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 approach we can compute the value $f(n)$ for arbitrary $n \in \mathbb{N}$. Note that this idea resembles the construction of natural numbers from the empty set or the principle of mathematical induction.

An example of such equation is the definition of the factorial function:

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

We see that the value of $f(n+1)$ depends on both n and the value of $f(n)$.

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

$$f(n+1) = a \cdot f(n) + b, \tag{1.6}$$

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

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

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

Thus for all n we have

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

This is also the relation for the so-called Malthusian population growth model, which is based on the assumption that during a given time interval the populations grows with a constant ratio a to the state before the interval.

We will prove a general result for first order equations, which are similar to linear, but allow varying coefficients of a and b,

$$f(n+1) = a_n \cdot f(n) + b_n. \tag{1.7}$$

First let us think about what such equations can describe.

Linear difference equation (1.6) can we nicely interpret 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.

Do not be afraid of the seemingly difficult calculations in the following result. It is a typical example of technical mathematical statement for which it is hard to "guess" precisely how it should be formulated. On the other hand, it is then a simple exercise on the properties of scalars and mathematical induction to prove it. Really interesting are then the corollaries, see 1.11 later.

In the formulation we use along with the usual notation for sum $\sum$ the similar notation for the product $\prod$. In the rest of the text we will also use the convention that when the index set is empty, then the sum is zero and that the product is one.

Solution. Let S denote the sum Michael has to pay per month. After first month Michael repays S, part of it is a repayment of the loan, part of it repays the interest. Let d_k stand for the loan after k months. After first month d_k is

$$d_1 = 30000 - S \; + \frac{0,06}{12} \cdot 30000.$$

In general, after k-th month we have

$$d_k = d_{k-1} - S + \frac{0,06}{12} d_{k-1}. \tag{1.1}$$

Using the relation (1.9) is d_k given by

$$d_k = \left(1 + \frac{0,06}{12}\right)^k 30000 - \left[\left(1 + \frac{0,06}{12}\right)^k - 1\right]\left(\frac{12S}{0,06}\right).$$

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

$$S = 30000 \left(\frac{\frac{0,06}{12}}{1 - (1 + \frac{0,06}{12})^{-36}} \right) \doteq 912.7. \tag{1.2}$$

□

Note that the recurrence relation (‖1.1‖) 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.

1.32. Consider the case from the previous example. For how long would Michael have to pay, if he would like to repay €500 per month?

Solution. Setting $q = \left(1 + \frac{0,06}{12}\right) = 1.005$, $c = 30000$ the condition $d_k = 0$ gives the equation

$$q^k = \frac{200S}{200S - c},$$

by taking logarithms of both sides we obtain

$$k = \frac{\ln 200S - \ln(200S - c)}{\ln q},$$

which for $S = 500$ gives approximately $k = 71,5$, thus Michael would be paying for six years (and the last repayment would be less than €500). □

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

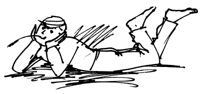
$$y_{n+1} = \frac{3y_n}{2} + 1,\; n \geq 1,\; y_1 = 1.$$

○

Linear recurrence can appear for instance in geometric problems:

1.10. Proposition. *General solution of the difference equation (1.7) of first order with the initial condition $f(0) = y_0$ is given by the formula*

$$f(n) = \left(\prod_{i=0}^{n-1} a_i\right) y_0 + \sum_{j=0}^{n-2}\left(\prod_{i=j+1}^{n-1} a_i\right) b_j + b_{n-1}. \tag{1.8}$$



Proof. We will prove the proposition using mathematical induction. It clearly holds for $n = 1$ where it amounts directly to the definition $f(1) = a_0 y_0 + b_0$.

Assuming that the statement holds for some fixed n, we can easily compute:

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

as can be directly seen by multiplying out. □

Let us again note that for the proof we did not need anything about the scalars we used except for the properties of commutative ring.

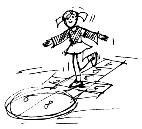
1.11. Corollary. *General solution of linear difference equation (1.6) with $a \neq 1$ and initial condition $f(0) = y_0$ is*

$$f(n) = a^n y_0 + \frac{1 - a^n}{1 - a} b. \tag{1.9}$$

Proof. If we set a_i and b_i to be constants and use the general formula (1.8) we obtain

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

For evaluating the sum of products in the second summand we need to observe that these are expressions $(1 + a + \cdots + a^{n-1})b$. The sum of this geometric progression can be computed using the formula $1 - a^n = (1 - a)(1 + a + \cdots + a^{n-1})$, and that yields the required result. □



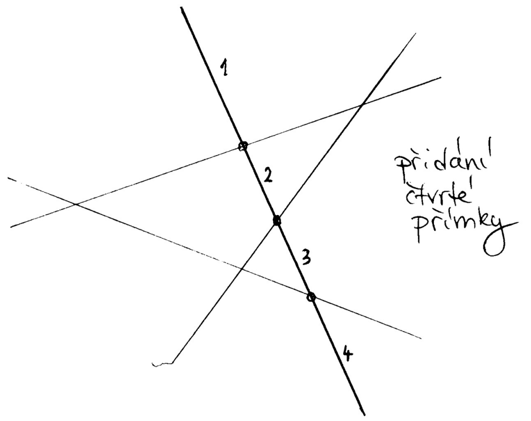
Note that for calculating the sum of a geometric progression we required the existence of the inverse element for non-zero scalars. We could not do that with integers only. Thus the last result holds for field of scalars and we can thus use it for linear difference equations where the coefficients a, b and the initial condition $f(0) = y_0$ are rational, real or complex numbers; and also in the ring of remainder classes $\mathbb{Z}_k$ with prime k (we will define remainder classes in the paragraph 1.41).

It is noteworthy that the formula (1.9) actually holds even with the integer coefficients and initial condition. Then we know in advance that all $f(n)$ are integer, and integers are a subset of rational numbers. Thus our formula necessary 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

1.34. Suppose n lines divide the plane into areas, what is the maximal number of areas that can arise this way?

Solution. Let the number of areas be p_n. If there is no line in the plane, then the whole plane is an area, thus $p_0 = 1$. If there are n lines, then by adding $(n+1)$-st line increases the number of areas by the number of areas this new line interesects. If no lines are parallel and no three lines intersect at the same point, the number of areas the $(n+1)$-st line crosses equals to 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 area, thus the new line crosses at most $n+1$ areas. Before adding the line, there was at most p_n areas (by the definition of p_n).



Thus we obtain the recurrence relation

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

from which we obtain an explicit formula for p_n either by applying the formula 1.10 or directly:

$$\begin{aligned} p_n &= p_{n-1} + n = p_{n-2} + (n-1) + n = \\ &= p_{n-3} + (n-2) + (n-1) + n = \cdots = p_0 + \sum_{i=1}^{n} i = \\ &= 1 + \frac{n(n+1)}{2} = \frac{n^2+n+2}{2} \end{aligned}$$

□

Recurrence relation can be more complex than first order. Let us list example of combinatorial problems, for whose solution a recurrence relation can be used.

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.

1.12. Nonlinear example. Let us return for a while to the first order equation (1.6) we have used for a very primitive population growth model which directly depends on the momentary population size p. On first sight it is clear that such model with $a > 1$ leads to a very rapid and unbounded growth.

A more realistic model has such 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 population grow by 5% for a time interval only for small p, we choose r to be 0, 05. For some limit value $p = K > 0$ the population does not grow and for even greater values it even decreases (since for instance the resources for the feeding of the population are limited, individuals in a big population are obstacles to each other etc).

Let us assume that exactly 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 variables p and y, which goes through the points $[0, r]$ (that is when $p = 0$ we have $y = r$) and $[K, 0]$ (which gives the second condition – when $p = K$ the population does not change). Thus we set

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

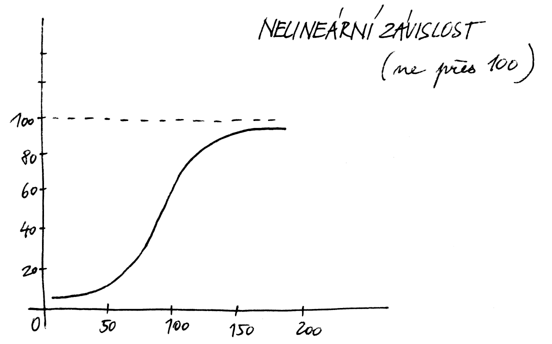
By setting $y = y_n$ and $p = p(n)$ we obtain

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

that is by multiplying out we obtain a difference equation first order (where the value of $p(n)$ is present as both first and second power).

$$p(n+1) = p(n)\big(1 - \frac{r}{K} p(n) + r\big). \tag{1.10}$$

Try to thing through the behaviour of this model for various values of r and K. On the picture we can see the values for parameters $r = 0, 05$ (that is, five percent growth in the ideal state), $K = 100$ (the resource limit the population to the size 100) and $p(0)$ are two individuals.



Note that the original almost exponential growth slows later down and the value approaches the desired limit of 100 individuals. For p close to one and K way greater than r the right side of the equation (1.10) is approximately $p(n)(1+r)$, that is the behaviour is similar to that of the Malthusian model. On the other hand, for p almost equal to K the right side of the equation is approximately $p(n)$. For initial value of p greater than K the values will decrease, for smaller than K they will grow, thus the system will basically oscillate about the value K.

1.35. 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

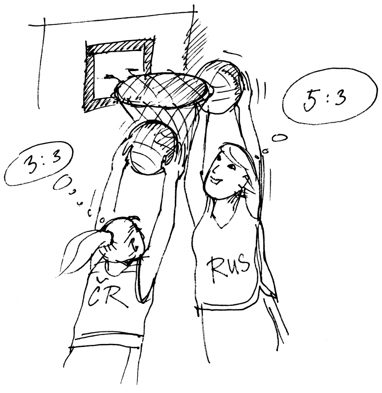
$$a_n = a_{n-1} + a_{n-2} + a_{n-3},$$

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, we can easily compute that $a_1 = 2$, $a_2 = 4$, $a_3 = 7$. 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 we have developed. Characteristic polynomial of the recurrence relation is $x^3 - x^2 - x - 1$ with one real and two complex roots, which we can express using the relations (||1.12||). □

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



Solution. If we denote $P_{(k,l)}$ 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:

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

(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 a goal and how worth it was (1, 2 or 3 points)). Using the symmetry of the problem, it clearly holds that $P_{(k,l)} = P_{(l,k)}$. Further

4. Probability

Let us have a look on another frequent example of scalar-valued functions – the observed values are often known neither explicitly by a formula nor implicitly by some description. They are a result of some randomness and we try to describe the *probability* of some outcome happening.

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

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

If we describe mathematical model of such throwing with a "fair" dice, we expect and thus also require that every side occurs with the same frequency. In words, we say that "every in advance chose side occurs with the probability $\frac{1}{6}$".

But if you try to manufacture with a knife such dice from wood, you probably observe that the relative frequencies will not be the same. In such situation, we can after a large number of tries count the relative frequencies of each label, and set these to be the probabilities in our mathematical description. But no matter how large the number of tries is, we cannot exclude the possibility that all the tries we did was some unlikely combination of the result and thus our model is not well chosen.

In the following part we will work with abstract mathematical description of probability in the simplest approach. The question how accurate or adequate for a specific real-world problem is out of the realms of mathematics. But that does not mean that such question are not for mathematician, quite the opposite (most likely in cooperation with some experts in the given area). Later we will return to probability and see it as a theory describing the behaviour of random processes or fully deterministic processes where not all determining parameters are known.

Mathematical statistics allows us to say how much can we expect that a given model corresponds to reality, or allows us to determine the parameters of the model in such way that the correspondence with the observations is high and simultaneously can estimate the reliability of the chosen model.

For both probability and statistics a complex mathematical theory is required, which we build over the course of few semesters.

On the example of our dice we can imagine it as follows: in the probability theory we work with the parameters p_i for the probabilities of individual sides and only require that these probabilities are non-negative and their sum is

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

When choosing specific values p_i for a specific dice in mathematical statistics we can then estimate the reliability of our mathematical model of the die.

Our humble goal for now is just to indicate how to abstractly capture the probabilistic considerations in formal mathematical objects. The following paragraphs are thus basically just exercises in simple operations with sets and combinatorics (that is, calculating the number of possibilities of satisfying the condition for finite sets).

we have for $k \geq 3$ that:

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

which along with the initial condition gives $P_{(0,0)} = 1$, $P_{(1,0)} = 1$, $P_{(2,0)} = 2$, $P_{(3,0)} = 4$, $P_{(1,1)} = 2$, $P_{(2,1)} = P_{(1,1)} + P_{(0,1)} + P_{(2,0)} = 5$, $P_{(2,2)} = P_{(0,2)} + P_{(1,2)} + P_{(2,1)} + P_{(2,0)} = 14$, gives

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

□

Remark. We see that the recurrence relation in this problem has a more complex form in comparison to the form we have dealt with in our theory and thus we cannot evaluate arbitrary number $P_{(k,l)}$ explicitly, we can evaluate it only by a subsequent computing from previous elements. Such an equation is called partial difference equations, since the elements of the equation are indexed by two independent variables (k, l).

We will talk more about recurrent formulas (difference equations) of higher orders with constant coefficients in chapter 3.

D. Probability

Let us state 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 whether the outcome of the experiment belongs to a subset of possible outcomes ("favourable outcomes"). The probability we are trying to determine then equals to the number of favourable outcomes divided by the total number of all outcomes. Classical probability can be used when we assume (know) that each of the possible outcome has the same probability of happening (for instance, fair dice throwing).

1.37. What is the probability that the roll of a dice results to a number greater than 4?

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

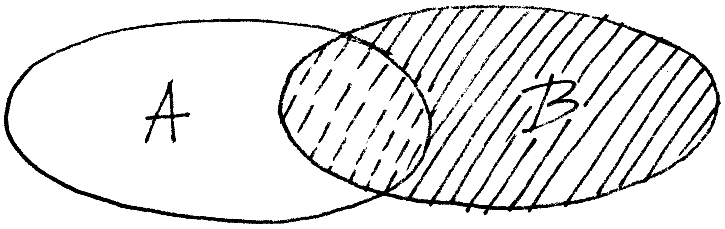
1.38. We randomly choose a group of five people from a group of eight men and four women. What is the probability that there are at least three women in the chosen group?

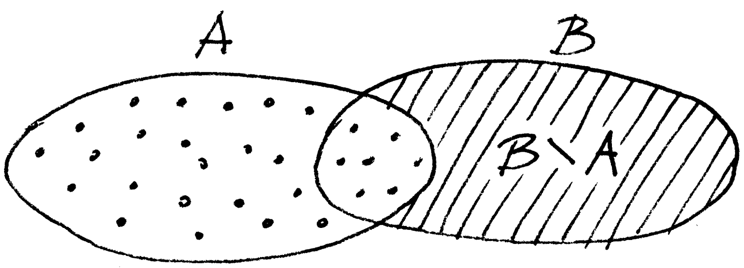
Solution. We compute the probability as a quotient of the number of favourable outcomes to the total number of outcomes. We divide the

1.14. Random events. We work with a non-empty fixed set Ω of all possible outcomes, which we call the *sample space*. For simplicity the set Ω is finite with elements $\omega_1, \dots, \omega_n$, corresponding to individual *possible outcomes*. Every subset $A \subset \Omega$ represent a possible *event*. The set of subsets $\mathcal{A}$ of the sample space is called the *set of events*, if

- $\Omega \in \mathcal{A}$ (the sample space is an event),
- if $A, B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$ (that is, for every two events their set difference is also an event),
- if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ (that is, for every two events their union is also an event).

PRAVDĚPODOBNOST





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

Clearly also the complement $A^c = \Omega \setminus A$ of an event A is an event, which we call the *opposite event* to the event A. The intersection of two events is again an event, since for every two subsets $A, B \subset \Omega$ holds

$$A \setminus (\Omega \setminus B) = A \cap B.$$

In words, the set of events can be also characterised as a system of subsets of (finite) sample space closed on intersection, union and set difference. Individual sets $A \in \mathcal{A}$ are called *random events* (with respect to $\mathcal{A}$).

For our dice throwing is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the set of events consists of all subsets of the set Ω. For instance the event $\{1, 3, 5\}$ is interpreted as "the result of the throw is an odd number".

Now for some terminology, which should remind of the connections with the description of real models:

- the whole sample space Ω is called *the sure event*, the empty subset $\emptyset \in \mathcal{A}$ is called *impossible event*,
- singleton subsets $\{\omega\} \subset \Omega$ are called *elementary events*, Pospisil definuje samotne prvky jako el.jevy.Jinde se to definuje i jeste jinak

favourable cases according to the number of men in the chose 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 we choose). 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 men from four, these two choices can be independently combined and thus using the rule of product we obtain the number of such groups). Thus the total number of groups with five people and at least three women is $c(12,5)=\binom{12}{5}$. The probability is then

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

□

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

1.39. 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. Favourable event is one (win), thus the probability is $1/2$ (a clearly wrong answer). □

Remark. In the previous exercise the basic condition of the usage 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 jackpot of €$25,000,000$ for choosing 5 correct numbers $1-50$. 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 equals $\frac{7k}{\binom{50}{5}}\doteq\frac{7k}{2118760}$.

1.40. There are $2n$ seats in a row in 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. In total there are $(2n!)$ of possible seatings, the number of seating satisfying the given condition is $2(n!)^2$: we have two ways for choosing the positions for men (thus also for women) – either all men sit on odd-numbered places (thus women sit on even-numbered places), or vice versa. Among these places, both men and women are

translators note: im really unsure about the english terminology regarding some of these terms...

- *intersection of events* A_i, $i\in I$, corresponds to the event $\cap_{i\in I}A_i$, *union of events* A_i, $i\in I$, corresponds to the event $\cup_{i\in I}A_i$,
- $A, B\in\mathcal{A}$ are *mutually exclusive*, if $A\cap B=\emptyset$,
- the event A has as a *corollary* the event B, if $A\subset B$,

Give an example of all listed terms for the sample space of dice rolling or analogously for coin throwing!

1.15. Definition. *Probability space* is a triple $(\Omega,\mathcal{A},P)$, where $\mathcal{A}$ is a set of events of (finite) sample space Ω, where there is a scalar function $P:\mathcal{A}\to\mathbb{R}$ with the following properties:

- P is non-negative, that is, $P(A)\geq 0$ for all events A,
- P is additive, that is, $P(A\cup B)=P(A)+P(B)$, whenever $A,B\in\mathcal{A}$ and $A\cap B=\emptyset$,
- the probability of the sure event is 1, that is $P(\Omega)=1$.

The function P is called *probabilistic* on the set of events $\mathcal{A}$.

Clearly an immediate corollary of our definitions is a list of simple yet useful propositions. For instance, for all events it holds that

$$P(A^c)=1-P(A).$$

Further, we can using mathematical induction extend the additivity to additivity for any number of mutually exclusive events $A_i\subset\Omega$, $i\in I$, that is,

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

whenever $A_i\cap A_j=\emptyset$, for all $i\neq j$, $i,j\in I$.

1.16. Definition. Let Ω be a finite simple space and let the set of events $\mathcal{A}$ be the set of all subsets of Ω. *Classical probability* is probabilistic space $(\Omega,\mathcal{A},P)$ with probabilistic function

$$P:\mathcal{A}\to\mathbb{R},\qquad P(A)=\frac{|A|}{|\Omega|},$$

where $|A|$ stands for the number of elements of the set $A\in\mathcal{A}$.

Clearly such given function is probabilistic, check by yourself that all the given axioms hold.

1.17. Summing probabilities. For mutually incompatible events is probability summing for the occurrence of at least one of them already incorporated into the definition of the probabilistic function. However, in general summing of probabilities for event occurrences is difficult. The problem is that whenever the events are mutually compatible, some of the elementary events are counted multiple times.

The simplest case to imagine is the one with two mutually compatible events A and B. Let us first consider classical probability, where it basically reduces just to counting elements in subsets. The probability of the occurrence of at least one of the events, that is, the probability of their union, is given by the formula

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)\tag{1.11}$$

since the elements that belong to both sets A and B were first counted twice and thus we have to subtract them once.

seated arbitrarily. The resulting probability is thus

$$p(n) = \frac{2(n!)^2}{(2n)!},\ p(2) \doteq 0,33,\ p(5) \doteq 0,0079,\ p(8) \doteq 0,00016.$$

□

1.41. Five persons entered 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 we have eight favourable outcomes, thus the probability is $\frac{1}{8^4}$ and finally 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 at the chosen floors), the probability is then (see 1.6 and 1.8)

$$\frac{v(5,8)}{V(5,8)} = \frac{8 \cdot 7 \cdots 4}{8^5} \doteq 0,2050781250.$$

□

1.42. Randomly choose a positive integer smaller than 10^5. What is the probability that it will consist only of digits 0, 1, 5 and it will be divisible by 5?

Solution. There are $2 \cdot 3^4 - 1$ numbers satisfying the conditions (except for the last digit, at every position we have three choices, if there are some 0 at the beginning of the number we ignore them but let them remain). There are $10^5 - 1$ positive integers smaller than 10^5, thus according to the classical probability we obtain that the probability is $\frac{2 \cdot 3^4 - 1}{10^5 - 1} \doteq 0.0016$. □

1.43. >From a sack with five white and five red balls we randomly draw three (we do not return the balls back to the sack). What is the probability that two of them are white and one is red?

Solution. Let us divide the event into a union of three disjoint events, according to in what turn we draw the red ball. Probability, that the red ball is drawn as third, second, or first, respectively, are : $\frac{1}{2} \cdot \frac{4}{9} \cdot \frac{5}{8}$, $\frac{1}{2} \cdot \frac{5}{9} \cdot \frac{1}{2}$, $\frac{1}{2} \cdot \frac{5}{9} \cdot \frac{1}{2}$. In total $\frac{5}{12}$.

Another solution. Consider the number of all possible triples of drawn balls (the balls of the same colour are indistinguishable), thus $\binom{10}{3}$. There are $\binom{5}{2} \cdot \binom{5}{1}$ of triples with exactly two white balls (two

The same result is obtained for general probabilistic function P on a set of events. Since $A \cap B$ and $A \setminus B$ are independent events,

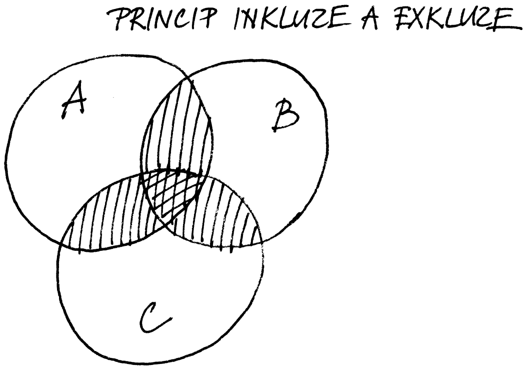
$$P(A) = P(A \setminus B) + P(A \cap B),$$

similarly for B, and we also have

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

If we express the probabilities of $P(A \setminus B)$ and $P(B \setminus A)$ in terms of $P(A)$, $P(B)$ and $P(A \cap B)$, we obtain the formula (1.11), now for the general case.

The following theorem is a direct reflection of the so-called combinatorial *inclusion-exclusion principle* into our finite probability and it says, how shall we deal with multiple elementary event counting in general case. It is probably a good example of mathematical theorem, where the hardest part is finding a good formulation. After that is done, we can say that the claim is (intuitively) obvious.



On the picture is the situation for three sets A, B, C for classical probability. If we just sum the probabilities of A, B and C, then the hatched areas represent elements that are present twice, and the double-hatched area represents those present thrice. Thus we subtract the hatched areas once, which leads to triple subtraction of the elements of double-hatched area, which we must therefore add once more.

In general, thanks to the additivity of the probability, we can imagine that we decompose every event as a union of elementary events (although the elementary events do not have to belong to the set of events in consideration). Then the probability of every event is given by the sum of probabilities of its elementary events. For expressing the probability that at least one of the events occurs we can sum the probabilities of A_i, then subtract those elementary events that are present twice (that is, in intersection of two A_i's). However, we might have subtracted some event too many times, notably in the case that an element was present in (at least) three A_i's. Thus we again add, and so on.

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

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

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 $\frac{\binom{5}{2}\cdot\binom{5}{1}}{\binom{10}{3}} \doteq \frac{5}{12}$. □

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

Solution. We will solve a 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 the classical probability. There are $P(5,5,6) = \frac{16!}{5!\cdot5!\cdot5!}$ 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 equals $P(5,5,5) = \frac{15!}{5!5!5!}$. Thus the probability is

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

□

Let us return to the dice throwing and try to describe the events of the sample space Ω that arise when we are throwing until a six is rolled, but no more than hundred times.

For a single roll the sample space consist of six numbers from one to six, this is classical probability. For whole series of rolls the sample space is much bigger – it consist of finite sequences of numbers from one to six, which have at most 100 elements and all numbers but the last one are from one to five, and either the last number is six or it has length exactly 100 and there is no number six in it. An event A can be for instance the subset "there were at most two rolls". All favourable elementary events are then

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

Using the classical probability for single dice rolls we derive the probability of the events in Ω. But it is not classical probability – if we were to derive the probability of the event A, it means that first roll is not six and the second is. Classical probability says

$$P(A) = \frac{5}{6}\cdot\frac{1}{6} = \frac{5}{36},$$

since the first roll is different from six with probability $1-\frac{1}{6}$, and the second roll is completely independent of the first one. Clearly, this does not amount to the quotient of all favourable cases to the size of the sample space.

Proof. In order to make the aforementioned ideas into a proof, we need to ensure that all the operations of adding subtracting are with coefficient one. Instead of doing that, we can give a more formal proof by mathematical induction over the number of events k, whose probabilities we are summing. Try to compare both approaches as they presented, it should help to clarify what it means to "prove" and what it means to "understand".

For $k=1$ the claim is obvious, the case $k=2$ is the same as the equality (1.11) which we have already proved (in general case).

Let us assume that the theorem holds for any number of events up to a definite $k \geq 1$. 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 (1.11) and the remaining event is considered to be the B :

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

This already resembles the formula for $k+1$ summed events, but still in the big sum there are missing the expressions containing A_{k+1} and the value for 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 paper long enough to write down all the expressions) we can check that this adds all the missing pieces. □

1.18. Inclusion-exclusion principle. A special case of the previous theorem is one of classical probability, where all finite subsets of the sample space are events and all elementary events have the same probability. In the formula from the previous theorem all the probabilities give the sizes of the subsets involved, up to a common factor $\frac{1}{n}$, where n is the number of elements of the sample space.

In this way we can extract from the theorem 1.17 the following claim for the size of a general finite set M and its subsets $A_1,\dots,A_k$. As usual we let $|M|$ denote the number of elements of the set M.

Of course that for every finite set M and its subspaces it holds that

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

Now we can use the previous theorem and express the size of the union on the right side, and we obtain the theorem that is usually called the *principle of inclusion-exclusion.*

$$\begin{aligned} &|M \setminus (\cup_{i=1}^{k} A_i)| = \\ &\qquad = |M| + \sum_{j=1}^{k}\Big((-1)^{j} \sum_{1\leq i_1<\cdots<i_j\leq k} |A_{i_1}\cap\cdots\cap A_{i_j}|\Big). \end{aligned}$$

Again it is easy to depict the theorem for two or three sets, see the picture before the theorem 1.17.

In general, we can say that after exactly $1 < k < 100$ rolls the experiment ends with probability $(\frac{5}{6})^{k-1} \cdot \frac{1}{6}$. Among all possibilities, it is most likely that it ends after the first roll.

Another example how to obtain events with different probability by dice throwing is to roll more dice and observe the sums. Let us think as follows: when rolling one dice every outcome has the same probability – $\frac{1}{6}$. When rolling two dice, every tuple (a, b) (a tuple of two integers from one to six in a given order) has the same probability $\frac{1}{36}$. If we ask for two fives, the probability is half of the probability for two different values without a prescribed order. When considering the sum of two dice throwing there is no contradiction with the classical probability concept as an event with a prescribed sum is a union of elementary outcomes that all have same probabilities $\frac{1}{36}$. For possible results of a sum of the two dice (in the upper row) we list the number of ways (in the lower row):

Sum	2	3	4	5	6	7	8	9	10	11	12
Number of ways	1	2	3	4	5	6	5	4	3	2	1

Similarly the probability for rolling three dice is $\frac{1}{216}$ (when assuming that order makes a difference). If we ask for the probability of some given sum to appear, we just need to determine in how many ways such a sum can appear and then sum the probabilities.

1.45. Inclusion-exclusion principle. 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 letter that was meant for him?

1.19. Independent events. Let us return for a while to the simple model of a fair dice. We are interested in possible dependencies among events.

For instance the probability that the events "an odd number is rolled" and "the result is at least three" occur simultaneously is $\frac{1}{3}$. That is the same as $\frac{1}{2} \cdot \frac{2}{3}$, the product of the probabilities of the events. This corresponds to the idea that the probability of simultaneous occurrence is given by the product of the particular probabilities. On the other hand, if we consider the mutually incompatible elements, for instance "even number occurs" and "odd number occurs", the probability of both of them occurring simultaneously zero, while the product of particular probabilities is nonzero. This corresponds to the idea that these two events must be the dependent, since the occurrence of the first one forbids the occurrence of the other one. Clearly, a weaker dependence can occur, for instance the event "odd number occurs" is a corollary of the event "number 3 occurs" and thus the probability of mutual occurrence is not given by the product.

For the probabilistic function P on an arbitrary set of events we say that the events A and B are *stochastically independent* if the following holds

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

Let us try the same approach with a dice and more events, for instance the events A "odd number occurs", B "the result is at least" and C "the result is at most 3". The probabilities are $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$, $P(C) = \frac{1}{2}$, $P(A \cap B \cap C) = \frac{1}{6} = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{2}$, but taking tuples we have for instance $P(A \cap C) = \frac{2}{3} \neq \frac{1}{2} \cdot \frac{1}{2}$.

In general, mutually independent sets are defined in this way:

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

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

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

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

>From there we can easily derive that by exchanging one or more event is a set of stochastically independent events we again obtain a set of stochastically independent sets.

Very often we need to compute the probability that at least one of the stochastically independent set of events occurs, that is, we want to compute $P(A_1 \cup \dots \cup A_k)$. In such situation we can use elementary properties of set operations, the so-called De Morgan laws,

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

and we obtain

$$\begin{aligned} P(A_1 \cup \dots \cup A_k) = 1 - P(A_1^c \cap \dots \cap A_k^c) = \\ = 1 - (1 - P(A_1)) \dots (1 - P(A_k)). \end{aligned} \tag{1.12}$$

Solution.

Let us compute the probability of the opposite event – no person receives the correct letter. The sample space corresponds to all possible orderings of six elements (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 ($i = 1, \dots, 6$) – so-called 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

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

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

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

and thus for the number d we obtain the relation

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

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

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

and the probability we were asked for is

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

□

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

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

as we see later, the sum converges to (approaches) the value $1/e$.

In a similar way the exercise ||1.153|| can be solved.

1.20. Conditional probability. We can reformulate the measure of dependency between two sets with the idea that we are investigating one of them under the condition that the other has occurred. For independent events, this does not have any impact. For instance, "what is the probability that in roll of two dice the result is twice 5, assuming that the sum of the results is 10?" We can formalise such approach as follows.

CONDITIONAL PROBABILITY

Definition. Let H be an event with non-zero probability in a set of events $\mathcal{A}$ in probability space $(\Omega, \mathcal{A}, P)$. *Conditional probability* $P(A|H)$ of the event $A \in \mathcal{A}$ assuming that H (the hypothesis) has occurred is defined by the formula

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

As it is obvious from the definition, the hypothesis H and the event A are 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

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

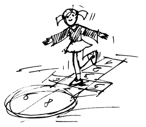
In words, this can be described as follows: the probability that three events occur at once can be computed by first computing the probability that the first occurs, then computing the probability that the second occurs under the assumption that the first has occurred, and then computing the probability that the third occurs under the assumption that both the first and the second have occurred.

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

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

Really, thanks to the assumption all the probabilities of the intersections, which are taken as the hypotheses, are non-zero. By simplifying the expression we obtain both on the left and on the right side of the equation the probability of the event corresponding to the intersection of $A_1, \dots, A_k$.

1.21. Geometric probability.



In practical problems we often encounter much more complicated models, where the sample space is not a finite set. At this moment, we do not have even basic tools for generalising probability to infinite sets at our disposal, but we can give at least a simple illustration.

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

1.46. Approximately 1200 persons die per year at the roads of Czech Republic. Determine the probability that some person of a chose 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 and that 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 does not die in a year is $(1 - \frac{12}{10^5})$. The probability that he 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 opposite event, that is, some of the chosen people dies, is then

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

□

Remark. Model we have used in the previous exercise to describe the given situation is just approximate. The complication is in the condition that every person in the sample has the same probability of dying, which we have 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. Let us show one of the possible inaccuracies on 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 of the probability

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

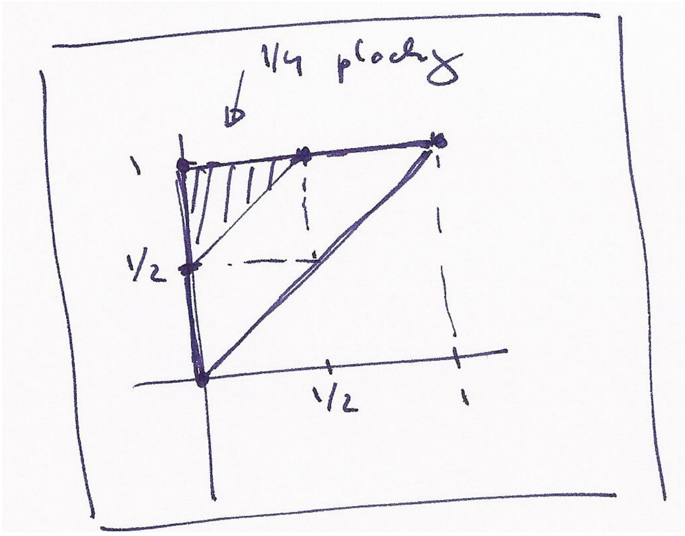
We see that both estimates are very close to each other.

The effort to use mathematical knowledge for winning in various gambling games is very old. Let us have a look on a very simple example.

1.47. Alex has a €2500 left over from organising a summer camp. Alex s is no fool – he added €50 from his savings and decided to go playing roulette. Alex s 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 bets twice the amount he betted in the previous (only if he has enough money, if not he ends the game even

Consider the plane $\mathbb{R}^2$ of tuples of real numbers and its subset Ω with known volume vol Ω. For example we can take the unit square. Events are represented by subsets $A \subset \Omega$ and for the event set $\mathcal{A}$ we consider some suitable system of subsets for which we can determine the volume. An event A then occurs if a randomly chosen point from Ω belongs to the subarea determined by A, otherwise the event does not occur.

Let us take for instance the problem where we randomly choose two values $a < b$ in the interval $[0, 1] \subset \mathbb{R}$. All values a and b are chosen with the same probability and the question is "what is the probability that the interval (a, b) has size at least one half?" The choice of points (a, b) is actually a choice of a point $[a, b]$ inside of the triangle Ω with border points $[0, 0]$, $[0, 1]$, $[1, 1]$ (see the picture).



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 him and his two friends. What is the probability that somebody gets at least a 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 inside of the triangle A bounded by the points $[0, \frac{1}{2}]$, $[0, 1]$, $[\frac{1}{2}, 1]$. Clearly we get that $P(A) = \frac{1}{4}$.

Try to answer on your own the question "what is the minimal prescribed length l such that the probability of choosing an interval of length at least l is one half?"

1.22. Monte Carlo methods. One of efficient computation methods for approximate values is the simulation of the probability by relative occurrence of a chosen event. For instance the well-known formula for the volume of a circle with given radius says that the volume of a unit circle is exactly the constant

$$\pi = 3,1415\ldots,$$

which expresses the ratio of the volume of the circle and the square of its radius. (Let us note that there is a fact we have not proven – why should the volume of a circle equal to a constant multiple of the square of its radius? We will be able to prove this mathematically after we learn how to do the so-called integration. Experimentally, we can verify this by the approach given bellow with squares of different size).

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 amount, he ends the game).

Solution. Let us first count how many times in a row can Alex s loose. If he begins with a bet of €10 , then for n bets he needs

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

As we can easily see, the number 2550 is of the form $10(2^n - 1)$ for $n = 8$. Alex s 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 amount (as soon as he has €5100, he ends). Thus in order for him to fail, he must lose eight times in a row. The probability of losing on one bet is 19/37, probability of losing eight times in a row is $(19/37)^8$ (as the bets are independent). 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. □

1.48. 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).

Let us now exercise the so-called "conditional" probability (see (1.20)).

1.49. What is the probability that when rolling two dice the sum is 7, if we know that neither of the rolls resulted in 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 independent. □

If we choose Ω to be unit square and A to be the intersection of Ω with a unit square (centred at the origin), then vol $A = \frac{1}{4}\pi$. Thus if we have a reliable generator of random numbers between zero and one and we compute relative frequencies how often the distance of the point $[a, b]$ (given by the generator) from the origin is smaller than one, that is, $a^2+b^2 < 1$, then the result (after a large number of attempts) approximates the number $\frac{1}{4}\pi$ pretty well.

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

5. Plane geometry

So far we have been using elementary notions from 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 again mappings, but this time we will consider only very special rules which to tuples of values (x, y) assign tuples $(w, z) = F(x, y)$. This part will also serve as a gentle introduction to the area of mathematics called *Linear algebra*, which we will deal with in subsequent three chapters.

1.23. Vector space $\mathbb{R}^2$. Let us view the "plane" as a set of tuples of real numbers $(x, y) \in \mathbb{R}^2$. We will call these tuples *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, the hold for our new vector addition too. In particular there exists so called *zero vector* $0 = (0, 0)$, whose addition to any vector v results again into the vector v. We are using the same symbol 0 for the vector and for its scalar coordinates on purpose — from the context it will be always clear which "zero" it should be.

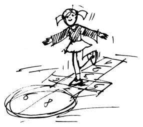
Next we define multiplication of vectors and scalars in such a way that for $a \in \mathbb{R}$ and $v = (x, y) \in \mathbb{R}^2$ we set

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

Usually we will omit the symbol $\cdot$ and just the juxtaposition of symbols $a\,v$ shall denote the scalar multiple of a vector. We can directly check other properties for scalar multiplication by a, b and addition of vectors u, v, for instance

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

where we are again using the same symbol plus for both vector addition and scalar addition.



These operation are easy to imagine if we consider the vectors v to be arrows going from the origin $0 = [0, 0]$ and ending at the position $[x, y]$ in the plane.

Such arrows can we compose — one right after another, and that corresponds to the vector addition. Multiplication by a scalar a corresponds to stretching the arrow to its a-multiple.

1.50. Michael has two mailboxes, one at gmail.com and the other at hotmail.com. His username is the same at both servers, but passwords are different (he does not remember which passwords corresponds to which server). When typing in the password for accessing his mailbox, he makes a typo with probability 5% (that is, if he wants to type in specific password, with probability 95% he types what he intended). Michael typed in at the server hotmail.com a 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" when typing in? (We assume that the username is always typed correctly)

Solution. Let A be the event that Michal typed in at hotmail.com a wrong password. This event is an 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 or not.

Thus we are looking for a conditional probability $P(A_1|A)$ which is according to the formula for conditional probability:

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

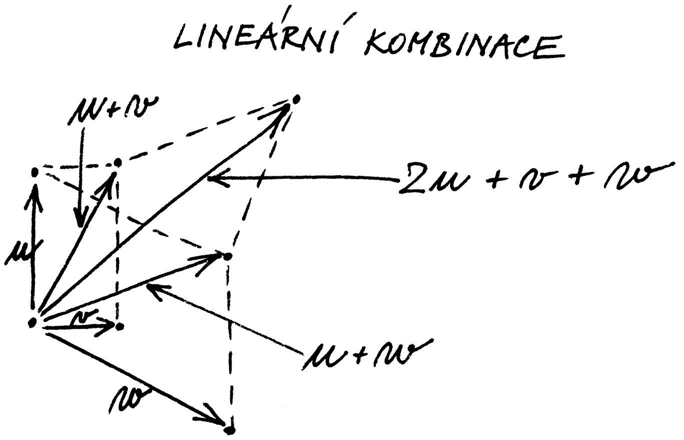
thus we just need to determine the probabilities $P(A_1)$ and $P(A_2)$. The event A_1 is a conjunction (intersection) of two independent events: Michael wanted to type in a correct password and Michael mistyped. According to the problem statement, the probability of the first event is $1/2$ and the probability of the second event is $1/20$, in total $P(A_1) = \frac{1}{2} \cdot \frac{1}{20} = \frac{1}{40}$ (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 $P(A) = P(A_1) + P(A_2) = \frac{1}{40} + \frac{1}{2} = \frac{21}{40}$, and we can evaluate

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

□

The method of *geometric probability* can be used in the case that the given sample space consists of infinitely many elementary events, which altogether fill some area of a line, space (where we can determine length, volume, …). We assume that the probability, that elementary event of a given sub-area happens, is equal to the ratio of the volume of the subarea to the volume of the sample space.

1.51. >From Edinburgh Waverly station trains depart every hour (in direction to Aberdeen) and from Aberdeen to Edinburgh they also comeevery 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



Now we are able to do a very important step: if we remember two important vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$, then every vector can be obtained as

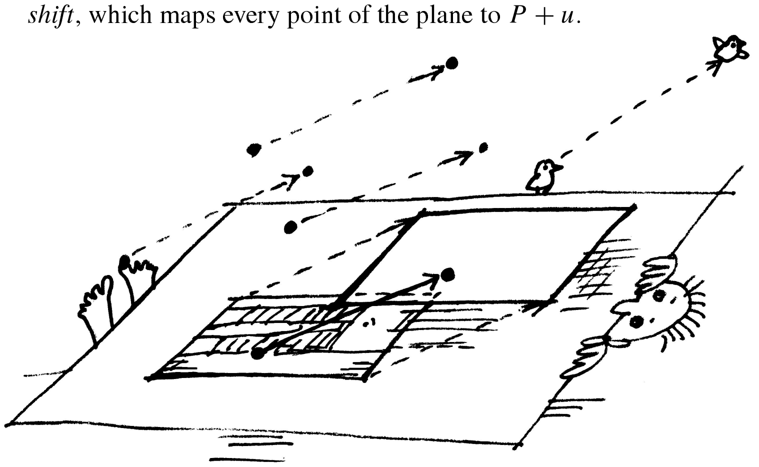
$$u = (x, y) = x\, e_1 + y\, e_2.$$

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

However, if we choose other two vectors u, v such that neither of them is a multiple of the other, that is a different basis of $\mathbb{R}^2$, we can do the same. Linear combination $w = x\,u + y\,v$ gives us for all distinct tuples (x, y) exactly all vectors w in the plane.

Finally, we can consider the vectors to be the arrows in the abstract position, that is if we forget the identification of the points in the plane with the tuples of numbers. The only fact that remains is that all the arrows are "fixed" in the point 0 which is also the zero vector. Operations of addition and scalar multiplication remain, and only through the choice of the base e_1, e_2 we identify our plane of arrows with $\mathbb{R}^2$.

1.24. Affine plane. If we fix some vector $u \in \mathbb{R}^2$, we can add it (that is, compose it with other vectors as an arrow) to any point $P = [x, y]$. Therefore with any fixed vector u we have defined *shift*, which maps every point of the plane to $P + u$.

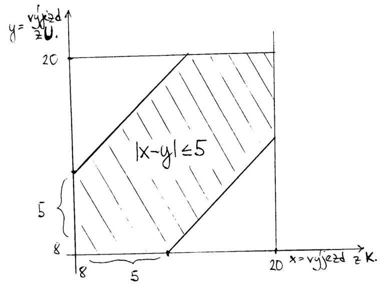


along the trail. After visiting an Edinburgh pub John, who lives in Aberdeen, takes train home and falls asleep at the departure. During the trip from Edinburgh to Aberdeen he randomly sticks his head out of the train for five seconds, in the space where the trains ride in the other direction. What is the probability that he loses his head? (We assume that there are no other trains here.)

Solution. The mutual speed of the oncoming trains is $40\,m/s$, the oncoming train passes John's window for two and a half seconds. The sample space of all outcomes is thus the interval $\langle 0, 7200\,s\rangle$. During John's trip two trains pass by John's window in the opposite direction and any overlap of the their $2.5\,s$ passing time interval with the $5\,s$ 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 $7.5\,s$ somewhere in the sample space. For two trains, it's double this amount. Thus the probability of losing the head is $15/7200 \doteq 0.002$. □

1.52. In one of the countries in the world, once a day between eight a.m. and eight p.m. a bus randomly departs from town A to town B. Once a day in the same time interval another bus departs in the other direction. The trip in either direction takes five hours. What is the probability that the buses meet, if they use the same trail?

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 in the square the are of "favourable events". The are of the remaining part is easier to compute, since it is an union of two right-angled isosceles triangles with legs of length 7. Thus in total it is 49, the area of the "favourable part" is $144 - 49 = 95$ and the probability is $p = \frac{95}{144} \doteq 0,66$.



□

1.53. 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.

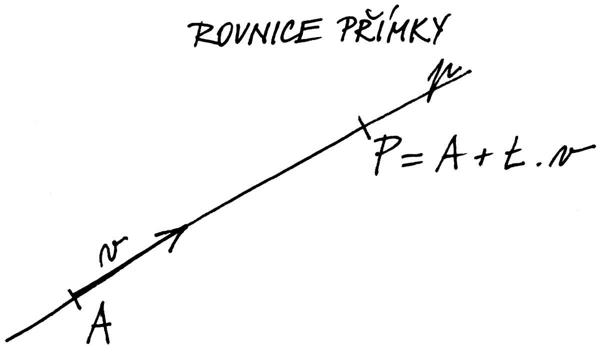
Solution. Random division of a rod into three parts is given by two points of the cut, x and y (we first cut the rod in the distance x from

Let us now completely forget the coordinates and see the whole plane as a set where are shifts take place. Such a set $A = \mathbb{R}^2$ can be imagined from the point of view of an observer, who observes from some fixed position (we can call that position for instance $O = [x_0, y_0] \in \mathbb{R}^2$). Suppose that the observer sees the plane as an infinite plate without any measurements and labels, and only knows what means shifting by any multiple of some vector $u \in \mathbb{R}^2$. Such a plane will be called "Affine plane".

In order to be able to see the "tuples of real numbers" around him, the observer must choose some fixed point E_1 which he will call the "point [1, 0]" and some other point E_2 which he will call the "point [0, 1]". In other words, he chooses a basis $e_1 = (1, 0)$, $e_2 = (0, 1)$ among the shifting vectors. To reach any point he will then just jump "a-times in the direction e_1" and then "b-times in the direction e_2" and the resulting point will be called the "point $[a, b]$". If he does it the usual way, it will not matter on the order of the operations, that is he can first jump b-times in the direction e_2 and after that in the direction e_1.

The thing we have described now is called the choice of *(affine) coordinate system in the plane*, the point O is its *origin*, and in general every point P of the plane is identified with the tuple of numbers $[a, b]$, which we will also denote as *shift* $P - O$.

>From now on we will work in fixed coordinates, that is with tuples of real number, but for better orientation we will denote vectors in parentheses instead of brackets (which we use for coordinates of points in the affine plane).



1.25. Lines in the plane. If our observer can shift by any multiple of a fixed vector, he also knows what is a *line*.

It is a subset $p \subset A$ in the plane, such that there exists point O and a non-zero vector v such that

$$p = \{P \in A;\ P - O = t \cdot v,\ t \in \mathbb{R}\}.$$

Let us now describe $P = P(t) \in p$ in the chosen coordinates with the choice $v = (\alpha, \beta)$:

$$x(t) = x_0 + \alpha \cdot t, \quad y(t) = y_0 + \beta \cdot t.$$

Since the vector $v = (\alpha, \beta)$ is non-zero, at least one of the numbers α, β has to be non-zero. Let us assume that for instance $\alpha \neq 0$, then we can eliminate t from the parametric equation for x and y and through a simple computation we obtain

$$-\beta x + \alpha y = -\beta x_0 + \alpha y_0.$$

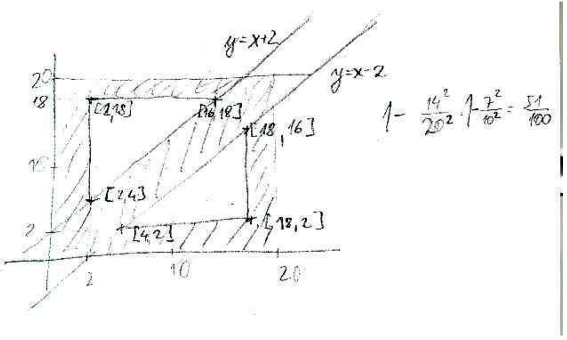
That is the general equation of the line

$$ax + by = c, \tag{1.13}$$

the origin, we do not move it and again cut it in the distance y from the origin). The sample space is thus 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 subarea O:

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

As we can clearly observe, this subarea has volume $\frac{51}{100}$ times the volume of the whole square.



y=x+2
y=x-2
20
18
10
2
2
10
20
[2,18]
[16,18]
[18,16]
[2,4]
[4,2]
[18,2]
$1 - \frac{14^2}{20^2} \cdot 1 - \frac{7^2}{10^2} = \frac{51}{100}$

□

E. Plane geometry

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

1.54. Interpret multiplication by the imaginary unit i and taking the complex conjugate as a geometrical transformations in the plane.

Solution. Imaginary unit i corresponds to the point $(0, 1)$. Let us note that multiplying any number $z = a + i\,b$ by the imaginary unit gives result

$$i \cdot (a + i\,b) = -b + i\,a$$

which is under the interpretation in the plane just a rotation of the point z through the right angle in the positive sense, that is counterclockwise.

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

$$z = (a + i\,b) \mapsto (a - i\,b) = \bar{z}.$$

□

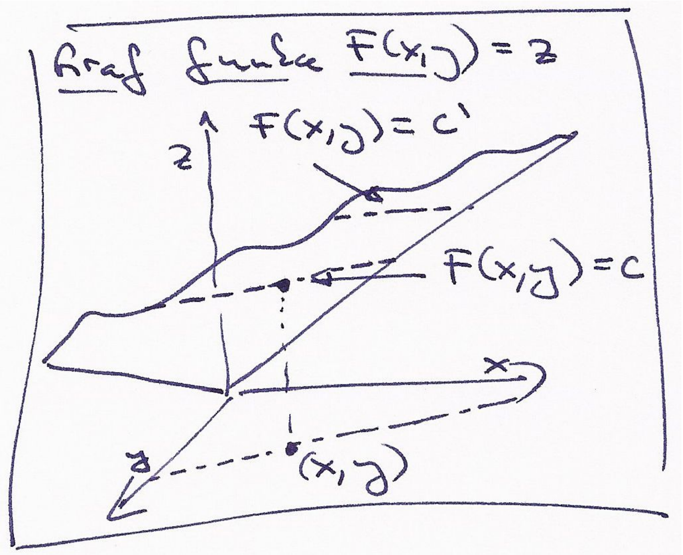
Now one well-known but nevertheless useful exercise.

1.55. Determine the sum of angles, which are between the vectors $(1, 1)$, $(2, 1)$ and $(3, 1)$ in the plane $\mathbb{R}^2$ (picture).

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 $3 + i$, and we are to find the sum of their arguments, which

with the following relation between the tuple of numbers $(a, b) = (-\beta, \alpha)$ and the direction vector of the line $v = (\alpha, \beta)$

$$a\alpha + b\beta = 0. \tag{1.14}$$



The expression on the right in the equation of the line (1.13) can we view as a scalar function F which depends on the points in the plane and with values in $\mathbb{R}$, the equation itself as a condition on its value. We shall see later that the vector (a, b) is in this case exactly the direction, in which F grows the fastest. For this reason will the direction perpendicular to (a, b) exactly the direction, in which our function F remains constant. The constant c then determines, which among all the parallel lines with that direction this equation corresponds to.

Let us now have two lines p and q, and ask about their intersection $p \cap q$. That is described as a point which satisfies the equations of both lines simultaneously. Let us write them like this

$$\begin{aligned} ax + by &= r \\ cx + dy &= s. \end{aligned} \tag{1.15}$$

Again we can view the left side as a mapping, which to every tuple of coordinates $[x, y]$ of point P in the plane assigns vector of values of two scalar functions F_1 and F_2 given by the left sides of the particular equations (1.15). Thus we can write our equations as a single relation $F(v) = w$, where F is a mapping which maps the vector v describing the position of any point in the plane (in our coordinates) to the vector given by the left side of the equations, and we demand that this mapping maps it to the specified value $w = (r, s)$.

1.26. Linear mappings and matrices. Mappings F with which we have worked with when describing the 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:

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

for all $a, b \in \mathbb{R}$, $v, w \in \mathbb{R}^2$. 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 is according to the de Moivre's formula 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 $\pi/2$ – thus the sum we looked for is exactly $\pi/2$. □

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

Solution. The vector $(-1, 3)$ gives the direction of the line p. Therefore the vector $(3, 1)$ is a normal to p and the characteristic equation of p is

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

for some $c \in \mathbb{R}$. We can determine this c by setting $x = 2$, $y = 1$ (the line p passes through point $[2, 1]$ with $t = 0$). Thus we obtain $c = -7$ and consequently the result $3x + y - 7 = 0$. □

1.57. We are given a line

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

Determine the characteristic equation of this line and the intersection with the line

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

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

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

We obtain the intersection of p with the line q by putting the points of q in parametric expression into the characteristic equation of p:

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

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

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

□

1.58. Determine the intersection of the lines

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

Solution. Let us first note that the direction of p is given by the vector $u_p = (1, -1)$ (any nonzero vector perpendicular to the vector $(1, 1)$ from the characteristic equation of p) and the direction of q is given by the vector $u_q = (2, 1)$. As the vector u_p is not a multiple of the vector v_q, we see that the lines have a nonempty intersection (they are not parallel). The point $[x, y]$ is the intersection if and only if its

with words — linear combination of vectors maps to the same linear combination of their images, that is linear mapping are those mappings which preserve linear combinations.

We have already encountered the same behaviour in the equation (1.13) 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 will write such mapping with the so called *matrices* and their multiplication. By matrix we mean a rectangular scheme of scalars, for instance

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{nebo} \quad v = \begin{pmatrix} x \\ y \end{pmatrix},$$

we speak of (square) matrix A and (column) vector v. The multiplication is defined as follows:

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

Similarly, we can instead of vector multiply from the right by another matrix B of the same dimension as A. We just apply the given formulas on individual columns of the matrix B and as a result we again obtain a square matrix.

We cannot multiply vector v from the right with matrix A because the number of scalars on the rows of v and the number of scalars in the columns of A differ. But we can write the vector w as a row of scalars (so called transposed vector) $w^T = (a\ b)$ and that we can multiply from the right with our matrix A or vector v already.

We can easily check the so-called associativity of multiplication (do it for general matrices A, B and a vector v in detail):

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

Of course that we can instead of vector v write any matrix C of correct dimension. In a similarly easy way can we see that distributivity also holds:

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

but the commutativity does not hold and there also exist "divisible zeros". For instance

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

We observe in particular that vector multiplication with a fixed matrix gives a linear mapping, and, in the other direction, using the values of a linear mapping F on two fixed vector of basis we obtain the whole corresponding linear mapping. Thus the points in the plane are in general images of the linear mapping F from a plane to a plane, lines are in general preimages of values of linear mappings from the plane to the real line $\mathbb{R}$. With matrices and vectors can we write the equations for lines and points as

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

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

Of course, in particular situations it does not have to be like this. For instance the intersection of two identical lines is the line itself (and the preimage of a specific value for such a linear mapping will be a whole line), the preimage of zero under the zero mapping coordinates satisfy the equation of p and there exist a real number t such that

$$x = -1 + 2t, \quad y = 2 + t.$$

If we put this into the equation of p, we obtain

$$(-1 + 2t) + (2 + t) - 4 = 0.$$

This equation is satisfied by $t = 1$, which gives the intersection with coordinates $x = 1$, $y = 3$. □

1.59. Find the characteristic equation of the line p, which goes through the point $[2, 3]$ and is parallel with the line $x - 3y + 2 = 0$, and the 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 $c \in \mathbb{R}$. The line q goes through the point $[2, 3]$. Therefore it must hold that

$$2 - 3 \cdot 3 + c = 0, \quad \text{tj.} \quad c = 7.$$

We can immediately give the parametric equation of the line q

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

□

1.60. Determine whether some of lines

$$p_1 : 2x + 3y - 4 = 0, \quad p_2 : x - y + 3 = 0, \quad p_3 : -2x + 2y = -6,$$

$$p_4 : -x - \tfrac{3}{2}\,y + 2 = 0, \quad p_5 : x = 2 + t,\ y = -2 - t,\ t \in \mathbb{R}$$

are parallel.

Solution. It is clear that

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

The characteristic equations thus describe the same line. The vector $(2, 3)$ is normal to the line p_1, for the line p_2 a vector is $(1, -1)$, for the line p_3 such a vector is $(-2, 2)$ and for the line p_5 such vector is $(1, 1)$ (perpendicular to the vector $(1, -1)$). The lines p_2 and p_3 are parallel (as the vectors perpendicular to them are multiples of each other). There are no more pairs of parallel lines, since the equations

$$x - y + 3 = 0, \quad -2x + 2y + 6 = 0$$

has clearly no solutions, the lines p_1 and p_4 form the only pair of identical lines. □

is the whole plane. The first case happens when on the left side of equations (1.15) there are the same expressions up to a scalar multiple (said in another way, the rows of the matrix A are the same up to a scalar multiple). In such a case either the intersection of the lines is empty (the lines are parallel but distinct) or it contains all points of the line (identical lines). This condition can be expressed by saying that the ratios a/c and b/d must be the same, that is

$$ad - bc = 0. \tag{1.16}$$

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

1.27. Determinant of matrix. The expression on the left in (1.16) is called *determinant* of the matrix A and we write for it

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

Our discussion can be now expressed as follows:

Proposition. *Determinant is a scalar function* $\det A$ *defined for all matrices A and equation $A \cdot v = u$ has a unique solution if and only if* $\det A \neq 0$.

It was necessary that we are working with the field of scalars — try to think it through. For instance it does not hold with integers in general. If we just compute the solution of the equations with integer coefficients (that is the matrix A has only integer inputs) the solution does not have to be integral in general.

1.28. Affine mappings. Let us now investigate how the matrix notation allows us to work with simple mappings in the affine plane. We have seen that matrix multiplication gives a linear mapping. Shifting in the affine plane $\mathbb{R}^2$ by a fixed vector $t = (r, s) \in \mathbb{R}^2$ can we also easily write in the matrix notation:

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

If we allow ourselves to add fixed vector to the result of a linear mapping then our expression will have the form

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

In this way we have described exactly all so-called *affine mappings of the plane* to itself.

Such mappings allow us recomputing of coordinates which arose by different choices of origins and bases of directions for shifting. What happens, if our observer from the paragraph 1.23 will observe the plane from a different point, or chooses different points E_1, E_2? Try to think through that when speaking about coordinates the difference will be exactly realised by affine mapping. Later we will see general reasons why that holds in any dimension.

1.61. 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 vector normal to q is the direction of p, we can directly write the result

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

□

1.62. 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 $a = b = 2$. □

1.63. 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. From the characteristic equations of p, q we obtain the vectors $(2, -1)$, $(1, 2)$ which are normal to them. The lines are parallel if and only if these vectors are multiples of each other, which is not the case. The intersection is found by solving the equations

$$2x - y - 5 = 0, \quad x + 2y - 5 = 0.$$

Expressing y from the first equation as $y = 2x - 5$ and putting it to the second equation we obtain

$$x + 2(2x - 5) - 5 = 0, \quad \text{tj.} \quad x = 3.$$

Then easily $y = 2 \cdot 3 - 5 = 1$. The intersection thus is $[3, 1]$. □

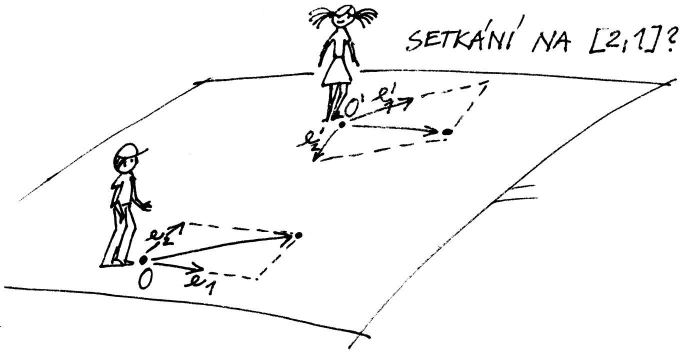
1.64. Consider the plane $\mathbb{R}^2$ with the standard coordinate system. A laser ray is sent from the origin $[0, 0]$ in the direction $(3, 1)$. It hits the mirror line p given by the equation

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

and then is reflected (the angle of rebound is the same as the angle of entry). At which point meets the ray the line q, given by

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

Solution. The angle between the line p and the direction of the ray is 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

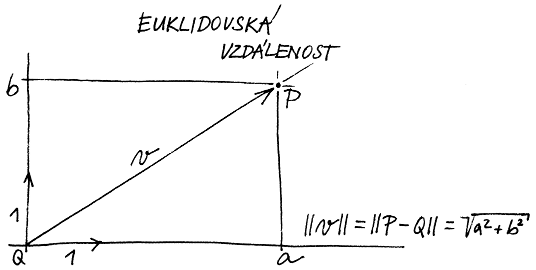


1.29. Euclidean plane. Let us now give our observer the ability to see and measure distance. For instance we can trust the common equation for the length of the vector $v = (a, b)$

$$\|v\| = \sqrt{a^2 + b^2}$$

in affine coordinates chosen by the observer. Immediately we can define notions as angle and rotation in the plane.

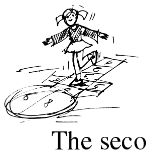
We can easily imagine it like this: our observer decides about some points E_1 and E_2 that they are at distance 1, and also decides that they are perpendicular. Distance in the direction of the coordinate axes are then given by the corresponding ratio, in general Euclid (Pythagorean) theorem is used. This leads to the equation given above.



Of course that our observer can work in a different manner. He can use some specific standard for real measurements of the distance of points P and Q in the plane and then say that exactly that is the length of the vector $Q - P$, which is necessary to shift from the P to Q. Then he picks some of the vectors which have size 1 and for instance using the triangle with sides of size 3, 4 and 5 constructs a perpendicular vector of size 1 and then continues as before.

Euclidean plane is an affine plane with a notion of distance as given above.

1.30. Angle between vectors. So-called trigonometric function



$(cos)\varphi$ — which we have already used in the discussion about complex numbers as points in the plane — is given by the value of the first coordinate of the unit vector whose angle with the vector $(1, 0)$ is φ.

The second coordinate of such vector is then clearly given by the real value $0 \le \sin\varphi \le 1$ satisfying

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

The angle between two vectors v and w can be in general described using coordinates $v = (v_x, v_y)$, $w = (w_x, w_y)$ like this:

$$\cos\varphi = \frac{v_x w_x + v_y w_y}{\|v\| \cdot \|w\|}.$$

the vector perpendicular to the line p). The ray meets the mirror at the point [6, 2], thus the rebounded ray has the equation

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

The intersection of the liven given by the rebounded ray with the line q is the point [4, 8], which lies out of the half-line given by the rebound ray ($t=-2$). Thus the rebound ray does not meet the line q. □

Remark. The reflection of a ray in three-dimensional space is studied in the exercise ‖3.53‖.

1.65. Line segment of length 1 started moving at noon with a constant speed 1 $mathrmms^{-1}$ in the direction (3, 2) from the point [−2, 0]. Another line segment of length of 1 has started moving from the point [5, −2] also at noon, but with double speed. Will they collide?

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

$$\begin{aligned} p &: [-2,0]+r(3,2),\\ q &: [5,-2]+s(-1,1). \end{aligned}$$

Characteristic equation of the line p is

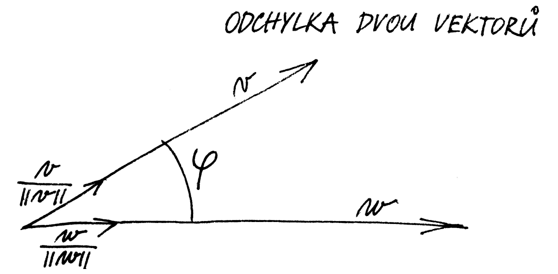
$$2x-3y+4=0.$$

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

Now let us try to choose a single parameter t for both lines so that the corresponding point describes the position at p and q of the first and second line segment respectively at the time t (more precisely, the position of the initial point of the line segment). At time 0 is the first line segment at the position [−2, 0], the second at the position [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 the corresponding parametrisations are:

$$\begin{aligned} p &: [-2,0]+\frac{t}{\sqrt{13}}(3,2),\\ q &: [5,-2]+t\sqrt{2}(-1,1). \end{aligned}$$

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

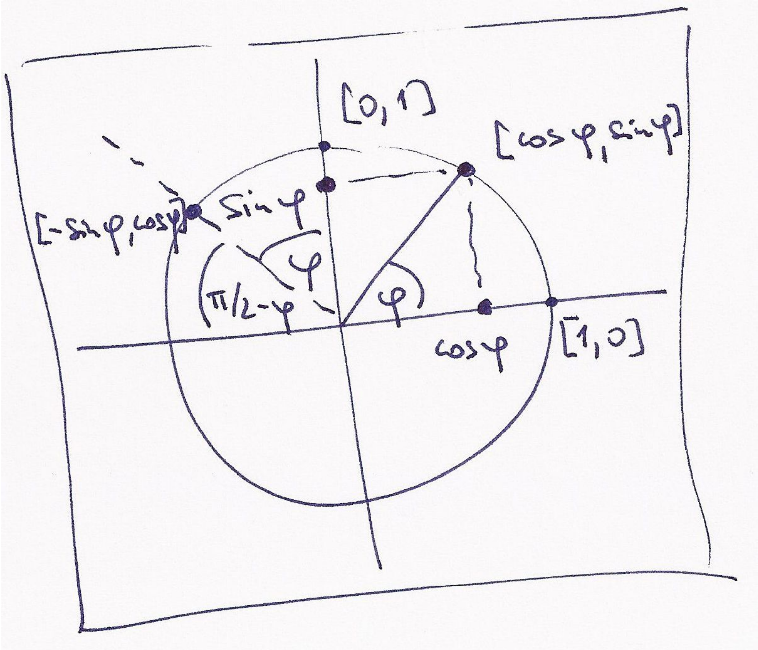


This relation can be easily check, as long as we believe that rotating the plane around the origin preserves angles. In such a case can we first multiply arbitrarily chosen vectors with suitable scalars such that we get vectors of length 1 (our equation clearly gives the same result after scalar multiplication). Then we can rotate the plane such that the first of our vectors coincides with the first basis vector (1, 0). Our equation gives then

$$\cos\varphi=\frac{w_x}{\|w\|},$$

which is just the definition of the function $\cos\varphi$.

1.31. Rotation around a point in the plane. 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 $(0,1)$.



We can see from the picture that rotating counter-clockwise through the angle ψ there are in the matrix the following columns:

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

The counter-clockwise direction is called the *positive direction*, the other direction is the negative direction. Therefore we obtain the claim:

of the second segment is already away and the segments do not collide. □

1.66. 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 $B = [26, 40]$. Does the ball fly towards the goal?

Solution. Due to the fact that the situation takes places in the first quadrant, it is sufficient to consider only the slopes of the vectors $\vec{FA}$, $(3, 4)$, $\vec{FB}$. If they form either increasing or decreasing sequence (in the order we have written them), the ball flies towards the goal. The sequence is $36/22$, $4/3$, $30/25$ which is a decreasing sequence, thus the ball flies towards the goal. □

1.67. Simplify $(A - B)^T \cdot 2C \cdot u$, while

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

Solution. By plugging in

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

and by matrix multiplication we obtain

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

□

1.68. 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. Let us remind that we are considering two-dimensional (square) matrices A and B. For any two 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 zero matrix, that is if and only if the matrices A and B commute. An example of such matrices are thus such pairs of matrices, which do not commute (the matrix of multiplication 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

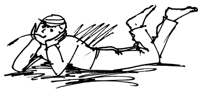
$$A \cdot B = \begin{pmatrix} 8 & 5 \\ 20 & 13 \end{pmatrix}, \quad B \cdot A = \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix}.$$

Analogously is for any pair of matrices A, B

Rotation matrix

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

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



Now, since we know how the matrix of the rotation in the plane looks like, we can check that rotation preserves distances and angles (defined by the previously given equation). Let us denote the image of a vector v as

$$v' = \begin{pmatrix} v'_x \\ v'_y \end{pmatrix} = R_\psi \cdot v = \begin{pmatrix} v_x \cos\psi - v_y \sin\psi \\ v_x \sin\psi + v_y \cos\psi \end{pmatrix},$$

and similarly $w' = R_\psi \cdot w$. We can easily check that it really holds that

$$\|v'\| = \|v\|$$
$$v'_x w'_x + v'_y w'_y = v_x w_x + v_y w_y.$$

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

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

The transposed vector $(R_\psi \cdot w)^T$ equals $w^T \cdot R_\psi^T$, 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 $R_\psi^T \cdot R_\psi = I$, the matrix I (sometimes we denote this matrix just as 1 and mean by this the unit in the ring of matrices) is the so-called *unit matrix*

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

This led 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. That 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 $P = O + w$, $P = [w_x, w_y]$ again using matrix, the equation can be expressed with a shift:

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

That means that

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

is satisfied if and only if $A{\cdot}B = B{\cdot}A$. In the second case is the answer exactly the same as in the first case. □

1.69. Decide, whether the mapping $F, G : \mathbb{R}^2 \to \mathbb{R}^2$ 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

$$F\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 7 & -3 \\ -2 & 5 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}, \quad G\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 2 & 2 \\ 4 & -9 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -4 \\ 3 \end{pmatrix}.$$

This implies that both mappings are affine. Let us remind that 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. □

1.70. Let us consider a regular hexagon $ABCDEF$ (the vertices are labelled in the positive direction) with centre at the point $S = [1, 0]$ and the vertex $A = [0, 2]$. 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:

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

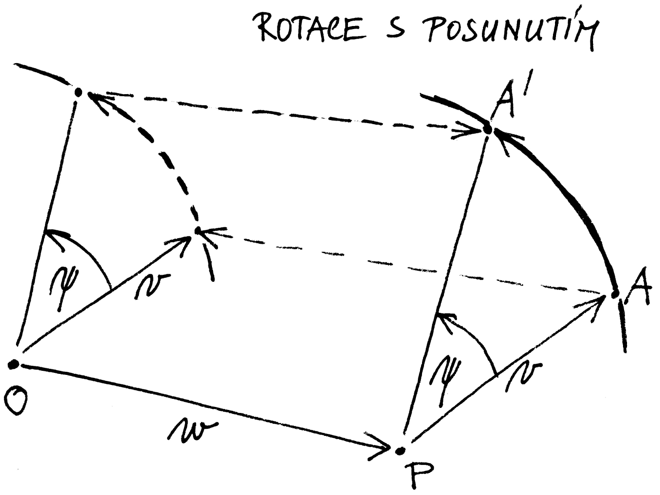
□

1.71. Determine the angle between two vectors

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

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

Solution. The angle φ we are looking for can be computed from the formula (1.36). 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



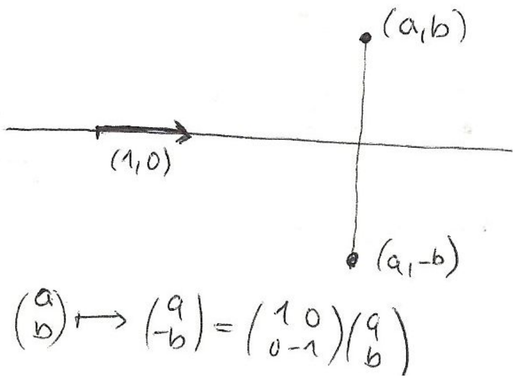
One just has to realise that instead of rotating around the given point P we can first shift P into the origin, then do the rotation and after that do the inverse shift, which takes the whole plane back where it should have been all the time. Let us calculate then:

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

1.32. Reflection. Another well-known example of mappings which preserve length is the so-called *reflection through a line*. Again it suffices to describe reflections through lines that go through the origin O, all other reflections can be derived using shifts and rotations.

Let us look for a matrix Z_ψ of reflection with respect to the line with the direction given by the unit vector v such that the angle between v and the vector $(1, 0)$ has value ψ. Let us first realise that

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



In general, we can rotate any line 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_O and return back with the rotation R_ψ.

(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 clearly zero. In our case we have to multiply by a negative number, which gives $\varphi = \pi$. □

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

Solution. The angle does not depended on the size of the given dodecagon. Let us choose the dodecagon inscribed in a circle with diameter 1. As in the previous exercise, we choose the coordinates of its vertices and then using a formula finish the computation that $\cos(\varphi) = \frac{1}{2\sqrt{2+\sqrt{3}}}$, that is $\varphi = 75°$.

Alternative solution. This problem can be solved via method of synthetic geometry only: let us 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 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 $180° - 105° = 75°$. □

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

Solution. Using the well-known formula for the size 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

$$\begin{aligned} |AB| &= ||A - B|| = \sqrt{(2-3)^2 + (2-0)^2} = \sqrt{5}, \\ |BC| &= ||B - C|| = \sqrt{(3-4)^2 + (0-3)^2} = \sqrt{10}, \\ |AC| &= ||A - C|| = \sqrt{(2-4)^2 + (2-3)^2} = \sqrt{5}. \end{aligned}$$

□

1.74. Let an equilateral triangle with vertices [1, 0] and [0, 1] which lies completely in the first quadrant be given. Determine 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). □

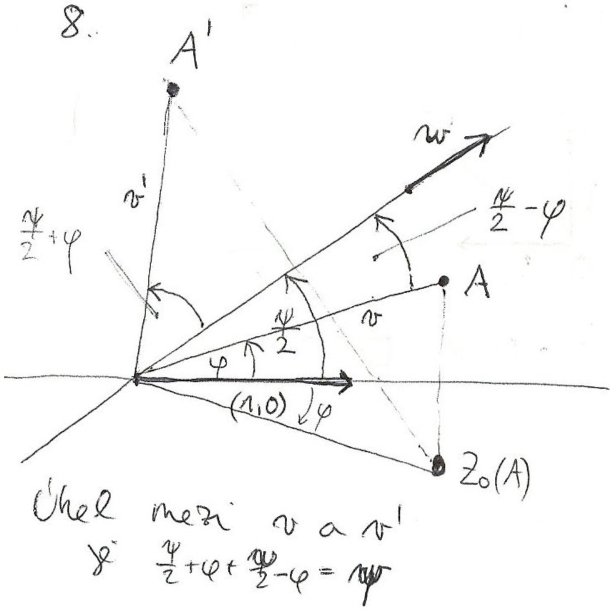
Therefore we can calculate (thanks to the associativity of matrix multiplication):

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

We have used the usual addition formulas for trigonometric functions. Let us also note that $Z_\psi \cdot Z_0$ is given:

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

This observation can be depicted and formulated as follows


8.
A'
w
v'
$\frac{\psi}{2} - \varphi$
$\frac{\psi}{2} + \varphi$
A
$\frac{\psi}{2}$
v
φ
(1,0)
φ
$Z_0(A)$
Úhel mezi v a v'
je $\frac{\psi}{2} + \varphi + \frac{\psi}{2} - \varphi = \psi$

Proposition. *Rotation through the angle ψ is obtained by two subsequent reflections in the directions that have the angle $\frac{1}{2}\psi$ between them.*

If we can prove the previous proposition purely via geometrical argumentation (try to be a "synthetic geometer"), we have just proved standard formulas for trigonometric functions of double angle.

The following recapitulation of previous ideas is somehow deeper (we can almost think that we can already prove some interesting mathematical result):

Mappings that preserve length

1.33. Theorem. *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 happen if and only if this mapping preserves length.

1.75. Determine the coordinates of the vertices of a triangle, which arises by rotating an equilateral triangle, whose two vertices are $A = [1, 1]$ and $B = [2, 3]$ and the third is in the half-plane given by the line AB and the point $S = [0, 0]$, by 60° in the positive direction around the point S.

Solution. The third vector of the triangle can be obtained for instance by rotating through 60° of one vertex around the other (in the correct direction). The points we are looking for then have coordinates $[-\frac{3}{2}\sqrt{3}, \sqrt{3} - \frac{1}{2}]$, $[\frac{1}{2} - \frac{1}{2}\sqrt{3}, \frac{1}{2}\sqrt{3} + \frac{1}{2}]$, $[1 - \frac{3}{2}\sqrt{3}, \sqrt{3} + \frac{3}{2}]$. □

1.76. 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 matrix we are looking for are

$$A = \pm \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix},$$

that is they are matrices of rotation through 30° or through 210°. □

1.77. 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

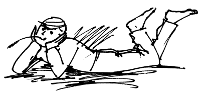
$$A \cdot A = \begin{pmatrix} \cos 2\varphi & -\sin 2\varphi \\ \sin 2\varphi & \cos 2\varphi \end{pmatrix}.$$

Let us 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

$$A^n = \begin{pmatrix} \cos n\varphi & -\sin n\varphi \\ \sin n\varphi & \cos n\varphi \end{pmatrix}, \quad n = 2, 3, \dots,$$

easier (we set $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$ etc.) □

Rotation is such 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. Let us first calculate, how a general matrix A may look like, if 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 it holds that

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

Since this equation is to hold for every x and y, the coefficients of the individual powers x^2, y^2 and xy on the left and right side of the equation must be equal. Thus we have calculated that the conditions put on the matrix R in the first part of the theorem we are proving are equivalent to the property than the given mapping preserves length.

Thanks to the relation $a^2 + c^2 = 1$ 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 size of the vector in the second column is one, and thus we have only two possible cases for the matrix R:

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

In the first case it is the rotation through the angle φ, in the second case it is the rotation composed with the reflection through the first coordinate axis. As we have seen in the previous proposition 1.31, every rotation corresponds to two reflections and the determinant of the matrix R is in these two cases really either one or minus one and distinguishes between these two cases. □

1.34. Area of a triangle. In the end of our little trip to the areas of geometry we will focus on the *area* of planar objects. For us, triangles will be sufficient. Every triangle is determined by a tuple of vectors v and w, which, if shifted 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 vol), which to two vectors assigns the number equal to the area vol $\Delta(v, w)$ of the triangle $\Delta(v, w)$ defined in the aforementioned way, where we for definiteness pick for P the origin (shift does not change the volume anyway).

We can see from the statement that the desired value is half of the area of the parallelogram spanned by the vectors v and w and is easy to calculate (using the well-known formula: base times corresponding height), or simply observe from the picture that the following holds

$$\begin{aligned} \operatorname{vol}\Delta(v + v', w) &= \operatorname{vol}\Delta(v, w) + \operatorname{vol}\Delta(v', w) \\ \operatorname{vol}\Delta(av, w) &= a \operatorname{vol}\Delta(v, w). \end{aligned}$$

1.78. Reflection. Find the matrix of reflection in the plane through the line $y = \sqrt{3}x$ (that is the matrix of the axial symmetry).

Solution. Draw a picture. In the basis given by the vectors $(1, \sqrt{3}), (-\sqrt{3}, 1)$ is the matrix of the reflection clearly $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Thus in the standard basis it is

$$\begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}^{-1}$$

The inverse matrix is in this case easy to find since the vectors in the columns are perpendicular, thus the matrix is (almost) orthogonal. We have

$$\begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}^{-1} = \frac{1}{4} \begin{pmatrix} 1 & \sqrt{3} \\ -\sqrt{3} & 1 \end{pmatrix}$$

By multiplying of the corresponding matrices we obtain the result $\frac{1}{2}\begin{pmatrix} -1 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}$. This result can be directly guessed from the picture (Redraw the pcture!). □

1.79. 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)

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix},$$

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

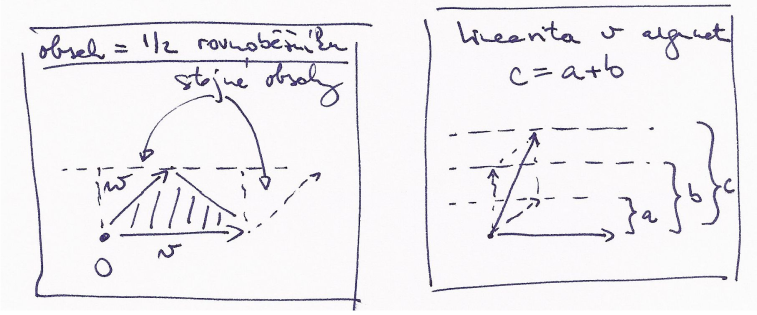
$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

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 counterclockwise). □

1.80. Parallelogram identity. Let us prove the so-called "parallelogram identity" for an illustration of our tools: If $u, v \in \mathbb{R}^2$, then:

$$2(\|u\|^2 + \|v\|^2) = \|u + v\|^2 + \|u - v\|^2.$$

That means, the sum of the squares of the diagonals of a parallelogram equals the sum of the squares of the lengths of the four sides of the parallelogram.



Finally we add to our problem formulation a condition

$$\operatorname{vol}\Delta(v, w) = -\operatorname{vol}\Delta(w, v),$$

which corresponds to the idea that we give a sign to the area, according to the order in which we are taking the vectors (that is, if we see the area from the top or from the bottom).

If we write the vectors v and w into the columns of a matrix A, then

$$A = (v, w) \mapsto \det A$$

satisfies all the three conditions we wanted. How many such mappings could there possibly be? Every vector can be expressed using two basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$ and by linearity is then every possibility for vol Δ uniquely determined by the value for these vectors. Since for area — in the same way as for determinant — is clearly $\operatorname{vol}\Delta(e_1, e_1) = \operatorname{vol}\Delta(e_2, e_2) = 0$ (due to the required antisymmetry), every such scalar function is necessarily determined by the value on the single tuple of arguments (e_1, e_2). Therefore all possibilities are equal up to a scalar multiple, which can be determined by the condition

$$\operatorname{vol}\Delta(e_1, e_2) = \frac{1}{2},$$

that is we are choosing *orientation* and *scale* through the choice of basis vectors and we want that the unit square has area equal to one.

Thus we see that the determinant gives the area of a parallelogram determined by the columns of the matrix A and the area of the triangle is thus one half of that.

1.35. Visibility in the plane. The previous description of the value for oriented area gives us 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 fixed 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 tell 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. For simplicity we can imagine that a line segment can be "seen" from the points to the right of it and cannot be seen from the points to left of it (this corresponds to the notion that object with bounded by line segments oriented counterclockwise has to the left of the line segments its interior, through which the segment cannot be seen).

Solution. Using both sides of the equation into the coordinates $u = (u_1, u_2)$, $v = (v_1, v_2)$ yields:

$$\begin{aligned}
&2(\|u\|^2 + \|v\|^2)\\
&= 2(u_1^2 + u_2^2 + v_1^2 + v_2^2)\\
&= u_1^2 + 2u_1v_1 + v_1^2 + u_2^2 + 2u_2v_2 + v_2^2 +\\
&\quad + u_1^2 - 2u_1v_1 + v_1^2 + u_2^2 - 2u_2v_2 + v_2^2\\
&= (u_1 + v_1)^2 + (u_2 + v_2)^2 + (u_1 - v_1)^2 + (u_2 - v_2)^2\\
&= \|u + v\|^2 + \|u - v\|^2.
\end{aligned}$$

□

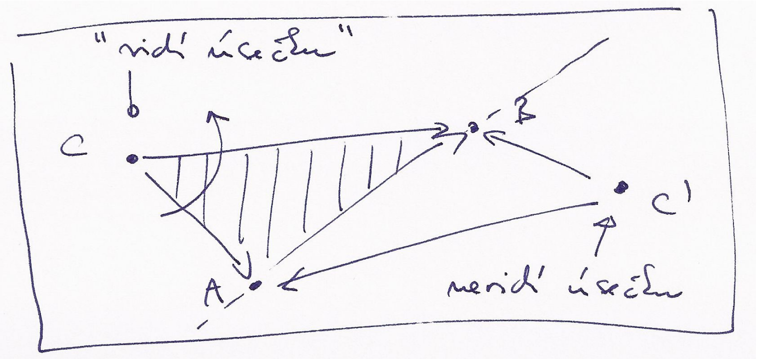
1.81. Show that by composing an odd number of point reflections in the plane yields again a point symmetry.

Solution. Point reflection in the plane across the point S is represented with the formula $X \mapsto S - (X - S)$, that is $X \mapsto 2S - X$. (The image of the point X in this reflection is obtained by summing the vector opposite to the vector $X - S$ and the vector S.) By repeated application of three point reflections across the points S, T and U respectively thus yields $X \mapsto 2S - X \mapsto 2T - (2S - X) \mapsto 2U - (2T - (2S - X)) = 2(U - T + S) - X$, that is $X \mapsto 2(U - T + S) - X$, which is a point reflection across the point $S - T + U$. Thus composition of any odd number of point reflection can be thus reduced to a composition of three point reflections, thus it is a point reflection (in principle, this is a proof by mathematical induction, try to formulate it by yourself). □

1.82. Construct $(2n + 1)$-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, $A_2, \dots, A_{2n+1}$ and the middle points of the sides (starting from the middle point of A_1A_2) by S_1, $S_2, \dots S_{2n+1}$. If we carry out the point reflections across the middle points, 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 $A_2, \dots, A_{2n+1}$ can be obtained by mapping the point A_1 in the point reflections across the points $S_1, \dots, S_{2n+1}$. □

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



We have the line segment AB and are given some point C. Let us now calculate the oriented area of the corresponding triangle determined by the vectors $A - C$ and $B - C$. If the point C is to the left of the line segment, then with the usual positive orientation (counter-clockwise) is the vector $A - C$ encountered sooner than the other vector $B - C$ and thus the resulting area (that is the value of the determinant of the matrix with these two vectors as columns) greater than zero. On the other hand, if the vectors are encountered in the other order, the resulting determinant value will be negative and thus we can say that the point is to the right of the segment.

The mentioned approach is really often used for testing the relative position in standard tasks in 2D graphics.

6. Relations and mappings

In the final part of the introductory chapter we will return to the formal description of mathematical structures, but we will try to illustrate them on examples we already know. Also, we can consider this part to be an exercise in formal approach to objects and concepts of mathematics.

1.36. Relations between sets. First we need to define *cartesian product* $A \times B$ of two sets A and B. It is the set of all ordered tuples (a, b) such that $a \in A$ and $b \in B$. *Binary relation* between two sets A and B is then a subset R of cartesian product $A \times B$.

Often we write $a \simeq_R b$ for expressing the fact that $(a, b) \in R$, that is that the points $a \in A$ and $b \in B$ are in relation R. *Domain of a relation* is the subset

$$D \subseteq A, \quad D = \{a \in A; \exists b \in B, (a, b) \in R\}.$$

In words, it is the set of elements a from the set A such that there exists an element b in B such that (a, b) belongs to the relation R. Shortly, the domain consists of such elements of A that have an image in B. Similarly *codomain of a relation* is the subset

$$I \subseteq B, \quad I = \{b \in B; \exists a \in A, (a, b) \in R\},$$

that is the elements of B that have a preimage in A.

Special case of a relation between sets is *mapping from a set A to the set B*. It is just for the case when every element of the domain of the relation is in relation with exactly one element of the codomain. Examples of mappings known to us are all scalar functions, where the domain of the mapping is a set of scalars, for instance the set of integers or reals. For mappings we usually use notation we have already been using when dealing with scalar functions. We write

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

Solution. We know that the area equals to 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 in the rows (preserving the order). □

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

Solution. Let us first denote the vertices (in the counter-clockwise direction)

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

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

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

$$d_2=\begin{vmatrix} 11-1 & 2-1 \\ 4-1 & 4-1 \end{vmatrix}=\begin{vmatrix} 10 & 1 \\ 3 & 3 \end{vmatrix},$$

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

$$S=\tfrac{d_1}{2}+\tfrac{d_2}{2}=\tfrac{5\cdot 3-10\cdot 0}{2}+\tfrac{10\cdot 3-1\cdot 3}{2}=\tfrac{15+27}{2}=21.$$

(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 $v=3$. □

1.85. Give the area of a meadow, which is determined on the area map by the points at quotas $[-7,1]$, $[-1,0]$, $[29,0]$, $[25,1]$, $[24,2]$ 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 for instance into for triangle 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, which gives the result

$$24+44\tfrac{1}{2}+13\tfrac{1}{2}+15=97.$$

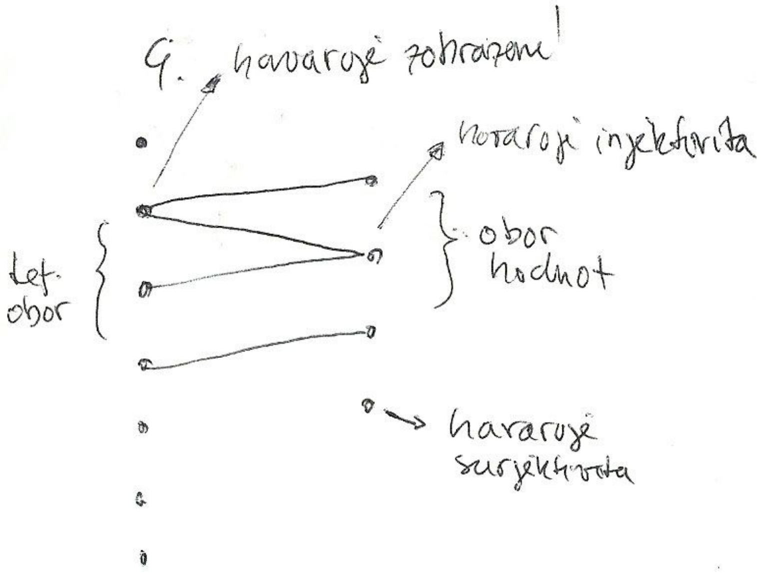
to express the fact that (a,b) belongs to a relation, and we say that b is the value of f at a. Furthermore we say that

- mapping f maps the set A to the set B if $D=A$,
- mapping f of the set A to the set B is *surjective* (or *onto*), if $D=A$ and $I=B$,
- 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

$$f\subseteq A\times B,\qquad f=\{(a,f(a));\,a\in A\}$$

is also known as *the graph of a function* f.



1.37. Composition of relations and functions. For mappings, the conception of composition is clear. 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

$$(g\circ f)(a)=g(f(a)).$$

It can be also expressed under the notation used for relation as

$$\begin{aligned} f&\subseteq A\times B, & f&=\{(a,f(a));\,a\in A\}\\ g&\subseteq B\times C, & g&=\{(b,g(b));\,b\in B\}\\ g\circ f&\subseteq A\times C, & g\circ f&=\{(a,g(f(a)));\,a\in A\}. \end{aligned}$$

The *composition of 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 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.

□

1.86. Determine the area of a triangle $A_2A_3A_{11}$, where $A_0A_1 \dots A_{11}$ are vertices of a regular dodecagon inscribed in a circle of radius 1.

Solution. The vertices of the dodecagon can be identified with twelfth roots of 1 in the complex plane. If we additionally choose $A_0 = 1$, we can then write $A_k = \cos(2k\pi/12) + i\sin(2k\pi/12)$. For the vertices of the investigated triangle it holds that $A_2 = \cos(\pi/3) + i\sin(\pi/3) = 1/2 + i\sqrt{3}/2$, $A_3 = \cos(\pi/2) + i\sin(\pi/2) = i$, $A_{11} = \cos(-\pi/6) + i\sin(-\pi/6) = \sqrt{3}/2 - i/2$, 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 is the area S equal to

$$S = \frac{1}{2}\begin{vmatrix} A_2 - A_{11} \\ A_3 - A_{11} \end{vmatrix} = \frac{1}{2}\begin{vmatrix} \frac{1}{2} - \frac{\sqrt{3}}{2} & \frac{1}{2} + \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{3}{2} \end{vmatrix} = \frac{3 - \sqrt{3}}{4}.$$

Since the determinant is non-negative, we could have omitted the absolute value for aestethical reasons. □

1.87. Which sides of the quadrilateral given by the vertices $[-2, -2]$, $[1, 4]$, $[3, 3]$ and $[2, 1]$ are visible from the position of the point $[3, \pi - 2]$?

Solution. It is a classical problem of the visibility of the sides of a convex polygon in the plane. In the first step we order the vertices such that their order corresponds the counter-clockwise direction. If we choose as the first vertex for instance the vertex $A = [-2, -2]$, the order of the remaining vertices is then $B = [2, 1]$, $C = [3, 3]$, $D = [1, 4]$. Let us first consider the side AB. It along with the point $X = [3, \pi - 2]$ determines the matrix

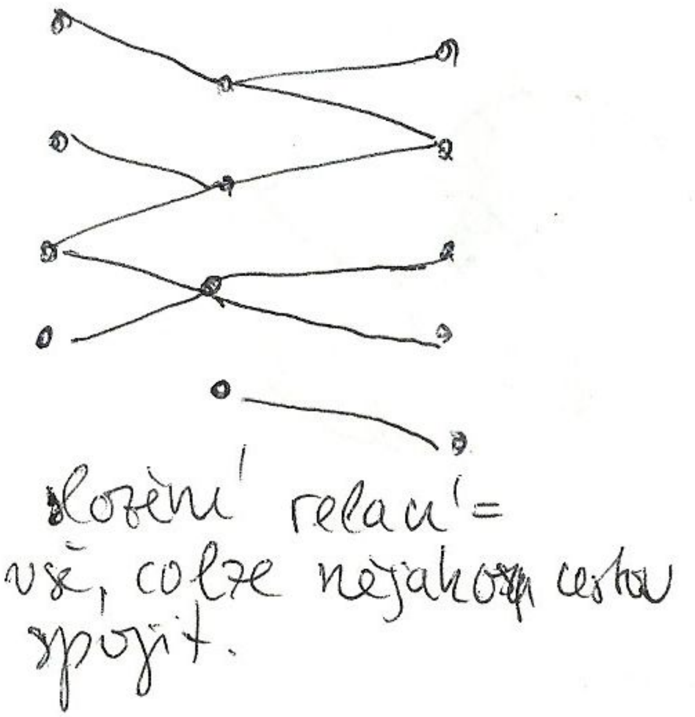
$$\begin{pmatrix} -2 - 3 & 2 - 3 \\ -2 - (\pi - 2) & 1 - (\pi - 2) \end{pmatrix}$$

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 $[3, \pi - 2]$, is then determined by the sign of the determinant

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

Negative value signifies that the side is visible. Let us note that it does not matter if we are considering the differences $A - X$ and $B - X$, or $X - A$ and $X - B$. But if we change the order of the columns, the corresponding side would be visible if and only if the determinant is positive.

For the side BC we analogically obtain



For every relation $R \subseteq A \times B$ we define the *inverse relation*

$$R^{-1} = \{(b, a); (a, b) \in R\} \subset B \times A.$$

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 case the inverse mapping is exactly the inverse relation.

Note that the composition of a mapping and its inverse mapping (if it exists) always leads to the identity mapping, but when dealing with relations it does not have to be so in general.

1.38. 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 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 $(a, c) \in R$.

Relation is called *equivalence* if it is reflexive, symmetric and transitive.

Relation is called *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 $a \leq b$.

Now it is good to realise that the relation $<$, that is "to be strictly smaller than", on real (rational, integer, natural) numbers is not an ordering, since it is not reflexive.

A good example of ordering is inclusion. Consider the set 2^A of all subsets of a finite set A (notation is a special example of the common notation B^A for the set of all mappings from A to

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

Thus this side is also visible. Only the sides CD and DA remain. For them we obtain

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

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

Thus from the point X are visible exactly the sides determined by the pairs of vertices $[-2,-2]$, $[2,1]$ and $[2,1]$, $[3,3]$. □

1.88. 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 let us „traditionally“ set

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

The sides BC and CD are clearly from the position of the point $[300,1]$ visible, on the other hand DE and EA cannot be seen. For the side AB let us determine

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

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

1.89. 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. Let us order the vertices in the positive direction, that is counter-clockwise: $[5,6]$, $[7,8]$, $[5,8]$. Using the corresponding determinants we can 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,

$$\begin{vmatrix} B-P \\ C-P \end{vmatrix} = \begin{vmatrix} 7 & 7 \\ 5 & 7 \end{vmatrix} >0, \qquad \begin{vmatrix} C-P \\ A-P \end{vmatrix} = \begin{vmatrix} 5 & 7 \\ 5 & 5 \end{vmatrix} <0,$$

$$\begin{vmatrix} A-P \\ B-P \end{vmatrix} = \begin{vmatrix} 5 & 5 \\ 7 & 7 \end{vmatrix} =0.$$

Since the last determinant is zero, we see that 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. □

B; the elements of the set 2^A are thus mappings from A to $\{0,1\}$ which "say" whether given element is in a given subset). We have a relation $\subseteq$ on the set 2^A given by the property "being a subset". Thus it is $X\subseteq Z$ 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 for every two elements $a,b\in A$ it holds that they are *comparable*, that means that either $a\leq b$ or $b\leq a$. Let us note that not all tuples (X,Y) of subsets of A are comparable in this sense. More precisely, if A contains more than element, there exist subsets X and Y where neither $X\subseteq Y$ nor $Y\subseteq X$.

Let us recall the recurrent definition of natural numbers $\mathbb{N}=\{0,1,2,3,\dots\}$, where

$$0=\emptyset,\qquad n+1=\{0,1,2,\dots,n\}.$$

On this set $\mathbb{N}$ we define a relation $\leq$ 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

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

In other words, the recurrent definition itself gives the relation $n\leq n+1$ and transitively then $n\leq k$ for all k obtained in this manner later.

1.39. Partitions of an equivalence. Every equivalence R on a set A gives also a *partition* of the set A, consisting of subsets of mutually equivalent sets, so called *equivalence classes*. For any $a\in A$ we consider the class (set) of elements, which are equivalent with a, that is

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

Often we will write for R_a simply $[a]a$, if it is clear from the context which equivalence we have in mind.

Clearly $R_a=R_b$ if $(a,b)\in R$, and every such equivalence class is therefore represented by any of its elements, so-called *representant*. Furthermore $R_a\cap R_b\neq\emptyset$ if and only if $R_a=R_b$, that is the equivalence classes are pairwise disjoint. Finally, $A=\cup_{a\in A}R_a$, that is the whole set A is partitioned to equivalence classes.

Another way of looking at things is that $[a]$ is seen as the element a "up to equivalence".

1.40. Construction of the integers and rational numbers. With natural numbers we can do addition and we know that adding zero to a number does not change it. We can also define subtraction, but the result does not always belong to the set $\mathbb{N}$.

The basic idea of construction of the integers from the natural numbers is to add to $\mathbb{N}$ these missing results. This can be done as follows: instead of result of subtraction, we will work with ordered tuples of numbers, which will represent the result. It remains just to define which such tuples are equivalent (with respect to the result of subtraction). The necessary relation is then:

$$(a,b)\sim(a',b')\iff a-b=a'-b'\iff a+b'=a'+b.$$

1.90. 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 $[2, 0]$.

Solution. First we need to determine the sides of the quadrilateral (the "correct" vertex order): $ABCD$. After computing the corresponding determinants as in previous exercises we see that only the side CB is visible. □

F. Mappings and relations

1.91. Determine whether the following relations over 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 in relation if they do not intersect.
iv) M is the set of lines in the plane, where two lines are in relation 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 number 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 it always holds that $C(n) < 10$).

Solution.

i) Yes. Let us check the three properties of equivalence:
    i) Reflexivity: for any real function f it holds that $f(0) = f(0)$.
    ii) Symmetry: if $f(0) = g(0)$, then also $g(0) = f(0)$.
    iii) Transitivity: if $f(0) = g(0)$ and $g(0) = h(0)$, then also $f(0) = h(0)$.
ii) No. The relation is not reflexive, since for instance for the function sin we have $\sin 0 \neq \sin 1$ and is not even transitive.
iii) No. The relation is not reflexive (every line intersects itself) and not transitive.
iv) Yes. The equivalence classes then correspond to unoriented directions in the plane.
v) No. The relation is not reflexive. $S(1) + S(1) = 2$.
vi) Yes.

□

1.92. We have a set $\{3, 4, 5, 6, 7\}$. Write explicitly the relations

Let us note that the expression in the middle equations cannot be realised in natural numbers, but the expression on the right can. We can easily 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 representants. For instance

$$[(a, b)] + [(c, d)] = [(a + c, b + d)],$$

which is clearly independent of the choice of representants.

It is always possible to choose representants $(a, 0)$ for natural numbers and representants $(0, a)$ for negative numbers — this is probably the simplest and clearest choice.

This simple example shows how important it is to be able to see the equivalence classes as a whole object and to concentrate on the properties of these objects, not on the formal description of their construction. However, the description is important in order to be able to check that such objects exist.

On integers we have all the properties of scalars (KG1)–(KG4) and (O1)–(O4), see the paragraphs 1.1 and 1.3. For multiplication the neutral element is one, but for all numbers a other than zero and one we are not able to find a number a^{-1} with the property $a \cdot a^{-1} = 1$, that means that for multiplication we are missing inverse elements.

Let us also note that the properties of the integral domain (ID), see 1.3. This means that if the product of two numbers equals zero, at least one of them has to be zero.

Thanks to the last stated property we can construct the rational numbers $\mathbb{Q}$ by adding all missing multiplicative inverses by a method analogous to the construction of $\mathbb{Z}$ from $\mathbb{N}$. On the set of all ordered tuples (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. Clearly this relation is a well-defined equivalence (think it through!) and rational numbers are then the equivalence classes. If we formally write p/q instead of tuples (p, q), we can define the operations of multiplication and addition by the well-known formulas.

1.41. Remainder classes. Another nice and simple example are the so-called 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. Again it is easy to see that using representants we can correctly define addition and multiplication for each $\mathbb{Z}_k$.

Theorem. *The remainder class $\mathbb{Z}_k$ is a commutative field of scalars (that is, the property (P) from the paragraph 1.3 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 clearly the result of multiplying the corresponding classes $[x] \cdot [y]$ is zero.

i) a divides b

ii) a divides b or b divides a

iii) a and b have a common divisor greater than one

○

1.93. 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 it is an equivalence relation. If it indeed is, describe geometrically the partitioning it determines.

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 left and right side. If $((a,b),(c,d)) \in R$ a $((c,d),(e,f)) \in R$, that is it holds that $b = d$ and $d = f$, we easily get that the transitivity condition $((a,b),(e,f)) \in R$, that is $b = f$, holds. The relation R is an equivalence relation, where the points in the plane are in relation 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. □

1.94. 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. Let us first realize 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 the correspond to subsets of this cartesian product, and of those there are exactly 2^{24}. □

1.95. 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. □

On the other hand, if x and k are relatively prime, then according to the so-called Bezout equality which we derive later (see **??**) natural numbers a and b satisfying

$$a\,x + b\,k = 1,$$

which for corresponding equivalence classes gives

$$[a] \cdot [x] + [0] = [a] \cdot [x] = [1]$$

and thus $[a]$ is the inverse element to $[x]$. □

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

$$R_a = \{(a,a), (b,b), (c,c), (d,d), (b,a), (b,c), (b,d)\},$$
$$R_b = \{(a,a), (b,b), (c,c), (d,d), (d,a), (a,d)\},$$
$$R_c = \{(a,a), (b,b), (c,c), (d,d), (a,b), (b,c), (b,d)\},$$
$$R_d = \{(a,a), (b,b), (c,c), (a,b), (a,c), (a,d), (b,c), (b,d), (c,d)\},$$
$$R_e = \{(a,a), (b,b), (c,c), (d,d), (a,b), (a,c), (a,d), (b,c), (b,d), (c,d)\}.$$

Solution. R_a is an ordering, which is not complete (for instance neither $(a,c) \notin R_a$ nor $(c,a) \notin R_a$). The relation R_b is not antisymmetrical (it is both $(a,d) \in R_b$ and $(d,a) \in R_b$), therefore it is not an ordering (it is an equivalence). The relations R_c and R_d are also not an ordering, since they are not transitive (for instance $(a,b), (b,c) \in R_c, R_d$, $(a,c) \notin R_c, R_d$) and also not reflexive ($(d,d) \notin R_c$, $(d,d) \notin R_d$). Relation R_e is a complete ordering (if we interpret $(a,b) \in R$ as $a \le b$, then $a \le b \le c \le d$). □

1.97. Determine whether the mapping f is injective (one-to-one) or surjective (onto), if

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

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) in is an injective mapping (both its coordinates, that is functions $y = 2x$ and $y = x^2 + 10$ are clearly increasing over $\mathbb{N}$) which is not surjective (for instance the tuple $(1,1)$ has no preimage). □

1.98. 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 injective?

Solution. To the element 1 we can assign arbitrarily one of the elements a, b, c. Similarly for the element 2 we have two possibilities. Thus according tho the (combinatorial) rule of product 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 $\{1, 2\}$. For the arbitrary mapping of the element 1 (three possibilities) we have injective mapping if and only if the element 2 gets mapped to a different element (two possibilities). Thus we see that the number of injective mappings of the set $\{1, 2\}$ to the set $\{a, b, c\}$ is 6. □

1.99. 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 us such a triple. Thus the number of injective mappings equals the number of ordered triples among four elements, that is $v(3, 4) = 4 \cdot 3 \cdot 2 = 24$. □

1.100. Determine the number of surjective mapping 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 $V(3, 4) = 3^4$, the number of non-surjective mappings (that is the number of mappings with one-element codomain) is three. Thus the number of mappings with 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

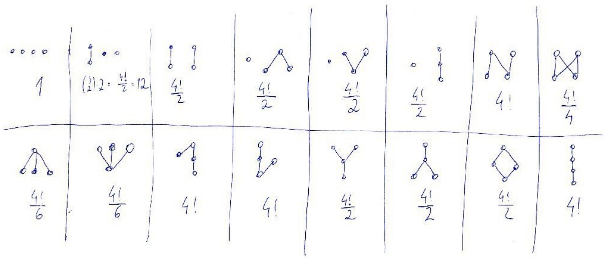
$$3^4 - \binom{3}{2}(2^4 - 2) - 3 = 36. \tag{1.3}$$

□

1.101. The Hasse diagram of ordering. The Hasse diagram of a give ordering $\prec$ over an n-element set M is a diagram with n vertices (every vertex corresponds to exactly one element of the set), and two vertices (elements) a, b are joined (with a more or less vertical) line (such that a is "lower" and b is "higher") if and only if b covers a, that is $a \prec b$ and there is no $c \in M$ such that $a \prec c$ and $c \prec b$.

1.102. Determine the number of ordering relations of a four-element set.

Solution. We will consider all possible Hasse diagrams of orderings over a four-element set M, and we will 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 picture:



In total, there are 219 orderings over a four-element set. □

1.103. Determine the number of ordering relations of the set $\{1, 2, 3, 4, 5\}$ such that exactly two pairs of element are incomparable.

○

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

Solution. Reflexive relations are exactly those, which contain both tuples $(1, 1)$, $(2, 2)$. By this we have excluded the relations

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

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) tuples, it must due to the symmetry condition contain also the other. If it contains none of these tuples, then it is clearly transitive. >From the total number of 16 relations over a two-element set have we thus chosen

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

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

1.105. Determine the number of equivalence relations over a set $\{1, 2, 3, 4\}$.

Solution. Equivalences can be enumerated by the sizes of their equivalence classes. For the sizes of equivalence classes over a four-element set we have these possibilities:

The sizes of equivalence classes	number of equivalences 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. □

Remark. In general, the number of partitions of a given n-element set is given by the *Bell number* B_{n+k}, for which a recurrence formula can be derived

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

1.106. How many relations are there over an n-element set?

Solution. 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}. □

1.107. How many reflexive relations are there over an n-element set?

Solution. Relation over the set M is reflexive if and only if it has the diagonal relation $\Delta_M = \{(a, a),\ \text{kde}\ a \in M\}$ as a subset. As for the rest of the $n^2 - n$ ordered tuples in the cartesian product $M \times M$ we have independent choice, whether the tuple belongs to the relation or not. In total we have 2^{n^2-n} different reflexive relations over an n-element set. □

1.108. How many symmetric relations are there over an n-element set?

Solution. 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, and 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, and every tuple (a, a), $a \in M$ either is in R or not. In total we have $2^{\binom{n}{2}+n}$ symmetric relation over an n-element set. □

1.109. How many anti-symmetric relations over an n-element set are there?

Solution. 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 $\{(b, a)\}$). 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 thus have $3^{\binom{n}{2}} 2^n$ anti-symmetric relations over an n-element set. □

In [**?**] we have defined remainder classes (also called residue classes) and we have shown that $\mathbb{Z}_p$ is a field for any prime p. On the other hand, events we are not used to when dealing with real or complex numbers occur in $\mathbb{Z}_p$.

1.110. Non-zero polynomial with zero values. Find a non-zero polynomial of one indeterminate with coefficients in $\mathbb{Z}_7$, that is an expression of the form $a_n x^n + \cdots + a_1 x + a_0$, $a_i \in \mathbb{Z}_7$, $a_n \neq 0$ such that it attains only zero values over the set $\mathbb{Z}_7$ (that is, if we set x to be equal to any of the elements of $\mathbb{Z}_7$ and evaluate, we always obtain zero).

Solution. For the construction of such polynomial we use the Fermat's little theorem which says that for any prime number p and number a, which is not divisible by p, we have:

$$a^{p-1} \equiv 1 (\text{mod } p).$$

Thus we can take for instance the polynomial $x^7 - x$ (the polynomial $x^6 - x$ is not zero for $x = 0$). □

G. Additional exercise for the whole chapter

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

Solution. 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 $t \cdot p^m = q^m$. Consider $\operatorname{ord}_r L$ and $\operatorname{ord}_r R$ and their divisibility by the number m. (L denotes the left-hand side of the equation, . . .). □

1.112. Determine

$$\left| \frac{(2+3i)\left(1+i\sqrt{3}\right)}{1-i\sqrt{3}} \right|.$$

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

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

□

1.113. Simplify the expression $\left(5\sqrt{3}+5i\right)^{12}$.

Solution. Taking powers one by one or doing an expansion using binomial theorem are in this case too much time-consuming. Let us rather write

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

and using the Moivre theorem we easily obtain

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

□

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

○

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

$$\bar{z} = \frac{\sqrt{3\sqrt{3}}}{2} - i\,\frac{3}{2}.$$

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

1.116. In the meeting there were six men. 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 $c\,(6,2) = \binom{6}{2} = 15$. □

1.117. Determine in how many ways a 4-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 4-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). □

1.118. 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

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

□

1.119. 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 $v\,(6,4) = 6 \cdot 5 \cdot 4 \cdot 3 = 360$. □

1.120. 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 $V(24,5) = 24^5$. □

1.121. 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

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

□

1.122. 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

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

□

1.123. 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

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

□

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

1.125. 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 $\binom{43}{3} = 12\,341$. □

1.126. 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 $\binom{12}{9} = 220$. □

1.127. 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, Oloumouc 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

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

□

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

i) Gouald a Bamba want to stand next to each other

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

iii) Gouald a 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 $6! - 2.5! = 720 - 240 = 480$.

□

1.129. 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?

○

1.130. How many anagrams of the word BAZILIKA 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 $2 \cdot 4! \cdot \frac{4!}{2!2!} = 288$. □

1.131. 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: $2^9(2^5 - 7) = 12800$. □

1.132. 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 a $6 \cdot 5^5$. □

1.133. 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 $x_1, \dots x_k$ are a solution to the equation if and only if the non-negative integers $y_i = x_i - 1$, $i = 1, \dots, k$ are a solution to the equation

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

Using ‖1.30‖, there are $\binom{n-1}{k-1}$ of them. □

1.134. There are n forts on a circle ($n \geq 3$), numbered in a row with numbers 1,..., n. In one moment of time each of the shoots at one of its neighbours (fort 1 neighbours 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

$$\begin{aligned} (K(n), K(n-1)) &= (K(n) - K(n-1), K(n-1)) = \\ &= (K(n-2), K(n-1)) = \cdots = 1. \end{aligned}$$

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

$$\begin{aligned} &(K(a), K(2a+1)) = (K(a), F(2)K(2a) + F(1)K(2a-1)) = \\ &= \ (K(a), F(3)K(2a-1) + F(2)K(2a-2) = \dots \\ &= \ (K(a), F(a+1)K(a+1) + F(a)K(a)) = \\ &= \ (K(a), F(a+1)) = (F(a+2) - F(a-2), F(a+1)) = \\ &= \ (F(a+2) - F(a+1) - F(a-2), F(a+1)) = \\ &= \ (F(a) - F(a-2), F(a+1)) = \\ &= \ (F(a-1), F(a+1)) = (F(a-1), F(a)) = 1 \\ &(K(a), K(2a-1)) = (K(a), F(2)K(2a-2) + F(1)K(2a-3)) = \end{aligned}$$

$$\begin{aligned}
&= (K(a), F(3)K(2a-3)+F(2)K(2a-4)) = \\
&= \cdots = (K(a), F(a)K(a)+F(a-1)K(a-1)) = \\
&= (K(a), F(a-1)) = (F(a+2)-F(a-2), F(a-1)) = \\
&= (F(a+2)-F(a), F(a-1)) = \\
&= (F(a+2)-F(a+1), F(a-1)) = (F(a), F(a-1)) = 1.
\end{aligned}$$

This proves the claim. □

1.135. 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)

$$\begin{aligned}
x_{n+1} &= 1,03(x_n) + 36000 + 1500 + \\
&\underbrace{0,03\cdot 3000\left(1+\frac{11}{12}+\cdots+\frac{1}{12}\right)}_{\text{interests from deposits this year}} + \\
&+ \underbrace{0,03\cdot\frac{2}{3}\cdot 1500}_{\text{interest from the state donation credited at this year}} = \\
&= 1,03(x_n) + 38115.
\end{aligned}$$

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

$$x_5 = 38115\left(\frac{(1,03)^4-1}{0,03}\right) + (1,03)^4\cdot 36585 + 1500 \doteq 202136.$$

□

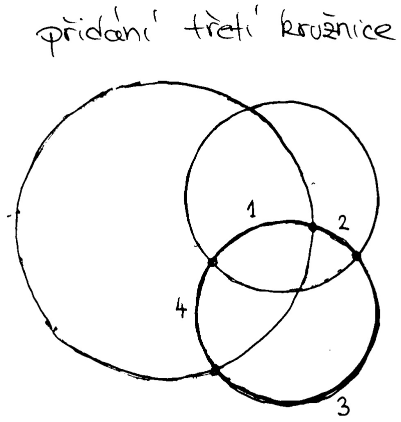
1.136. Remark. In reality, interests are computed according to the number of days the money is on the account. You should obtain a real bank statement of a building savings, determine its interest rates and try to compute the credited interests in a year. Compare the result with the sum that was credited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

1.137. 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)



Clearly $p_1 = 2$. Thus for p_n we obtain

$$\begin{aligned} p_n &= p_{n-1} + 2(n-1) = p_{n-2} + 2(n-2) + 2(n-1) = \dots \\ &= p_1 + \sum_{i=1}^{n-1} 2i = n^2 - n + 2. \end{aligned}$$

□

1.138. 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 (||1.34||) 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:

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

$$\begin{aligned} &= 1 + \frac{n(n+1)(2n+1)}{12} - \frac{n(n+1)}{4} + n = \\ &= \frac{n^3 + 6n + 5}{6}, \end{aligned}$$

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.

□

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

○

1.140. 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,

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

□

1.141. 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

$$\frac{\binom{26}{10}\cdot\binom{26}{6}}{\binom{52}{16}} \doteq 0,118.$$

□

1.142. 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

$$\frac{\binom{7}{2}\cdot 11}{\binom{18}{3}} \doteq 0,283.$$

□

1.143. 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

$$1 - \frac{\binom{96}{4}}{\binom{108}{4}} \doteq 0,380.$$

□

1.144. 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 $1/6$. □

1.145. 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. □

1.146. 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 $6 > 1$. □

1.147. 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 $1/2^n$. □

1.148. 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 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,

$$P = 0,4 \cdot (1 - 0,3) + (1 - 0,4) \cdot 0,3 = 0,46.$$

□

1.149. 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

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

□

1.150. 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 $2^2 00$ 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

$$kfrac\binom{200}{100}2^{200} = \frac{\frac{200!}{100!\cdot 100!}}{2^{200}} \doteq 0,056.$$

□

1.151. 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,

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

□

1.152. 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

$$n > \tfrac{\log_a 0,1}{\log_a 0,9}, \quad \text{kde } a > 1.$$

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

1.153. 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 $\binom{13.4}{2} = 1326$. 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 $4.51 - 6 = 198$ tuples and the probability is $\frac{198}{1326} \doteq 0,15$.
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 $1,5.10^{-6}$. 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 $\frac{32}{2598960} \doteq 1,2.10^{-5}$.
POKER: Four identical symbols – 13 choices (for every symbol one). The fifth card can be arbitrary, that is 48 choices. That makes $\frac{624}{2598960} \doteq 2,4.10^{-4}$.

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 $\frac{3744}{2598960} \doteq 1,4.10^{-3}$.
FLUSH: All five cards of the same colour means $4\binom{13}{5} = 5148$ choices and the probability is then $\frac{5148}{2598960} \doteq 2.10^{-3}$.
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 $9.4^5 = 9216$ 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 $3,5.10^{-3} + 2.10^{-3} + 1,4.10^{-3} + 2,4.10^{-4} = 7,14.10^{-3}$.

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 %.

□

1.154. 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 $\frac{315}{5040} = \frac{1}{16}$.

□

1.155. 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

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

□

1.156. 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 $2/6 = 1/3$. □

1.157. 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

$$P(A|H) = \tfrac{P(A\cap H)}{P(H)} = \tfrac{\frac{1}{36}}{\frac{4}{36}} = \tfrac{1}{4}.$$

Note that the sum 9 occurs when the first die is 3 and the second 6, the first is 4 and the second 6, the first is 5 and the second is 4, or the first is 6 and the second is 3. Of those four results (that have the same probability) only one is favourable to the event A. Since the probability of A is clearly $1/6 \neq 1/4$, the events are not mutually independent. □

1.158. 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

$$\tfrac{4}{32} \cdot \tfrac{3}{31} + \tfrac{28}{32} \cdot \tfrac{4}{31} = \tfrac{1}{8}.$$

□

1.159. 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 h stands for „girl“ considering the age of the children have the same probability. Choose random events

$$H_1 \text{ – familiy has a boy,} \quad A_1 \text{ – family has two boys.}$$

Compute $P\left(A_1|H_1\right)$.

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

$$P\left(A_1|H_1\right) = \frac{P(A_1 \cap H_1)}{P(H_1)} = \frac{P(A_1)}{P(H_1)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Further we have to determine whether the following holds:

$$P\left(A_2 \cap H_2\right) = P\left(A_2\right) \cdot P\left(H_2\right).$$

Again we just have to realize that exactly the elements kkk, kkh, khk, hkk of the set Ω, are favourable to the event A_2; to the event H_2 the elements kkh, khk, hkk, khh, hkh, hhk are favourable and to the event $A_2 \cap H_2$ the elements kkh, khk, hkk. Therefore

$$P\left(A_2 \cap H_2\right) = \frac{3}{8} = \frac{4}{8} \cdot \frac{6}{8} = P\left(A_2\right) \cdot P\left(H_2\right),$$

which means that the events A_2 and H_2 are independent. □

1.160. 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 $P\left(H^C\right)$ of the complementary event to the event H is 2^{-5} and that the probability of the event is the same as the probability $P\left(A^C\right)$ 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

$$P(A|H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{1}{2}}{1 - \left(\frac{1}{2}\right)^5} = \frac{16}{31}.$$

□

1.161. 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

$$\tfrac{9}{16} \cdot \tfrac{8}{15} \cdot \tfrac{7}{14} = 0,15.$$

□

1.162. 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

$$\tfrac{5}{10} \cdot \tfrac{5}{9} \cdot \tfrac{4}{8} \cdot \tfrac{3}{7} = \tfrac{5}{84}.$$

□

1.163. 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

$$\tfrac{28}{32} \cdot \tfrac{27}{31} \cdot \tfrac{26}{30} \cdot \tfrac{25}{29} \cdot \tfrac{24}{28} \cdot \tfrac{4}{27} \doteq 0,0723.$$

□

1.164. 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

$$P(A) = \tfrac{\operatorname{vol} A}{\operatorname{vol} \Omega} = \frac{\left(\frac{3}{7}\right)^2/2}{1} = \tfrac{9}{98}.$$

□

1.165. 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

$$\frac{\frac{d^2}{3^2\cdot 2}}{\frac{d^2}{2}} = \frac{1}{9}.$$

□

1.166. 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.

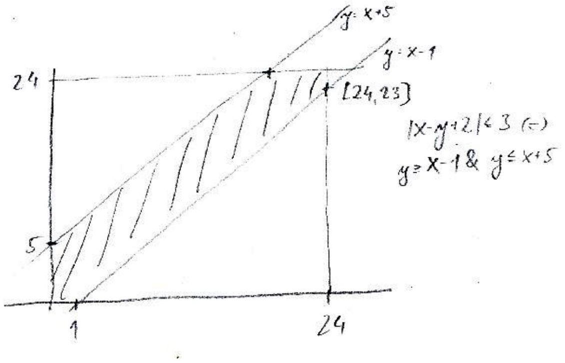
Solution. This exercise is for using the geometrical probability, where we are looking for the probability that the sum of the lengths of the first two parts is greater than one fourth of the length of the pole. We determine the probability of the complementary event, that is, the probability that if 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 $15/16$. □

1.167. 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 $|x - y| \leq 1/2$. 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. □

1.168. >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, $|(x+6) - (y+4)| \leq 3$. This condition determines an area with volume $24^2 - \frac{1}{2}(23^2 + 19^2)$ (see the figure) and the probability

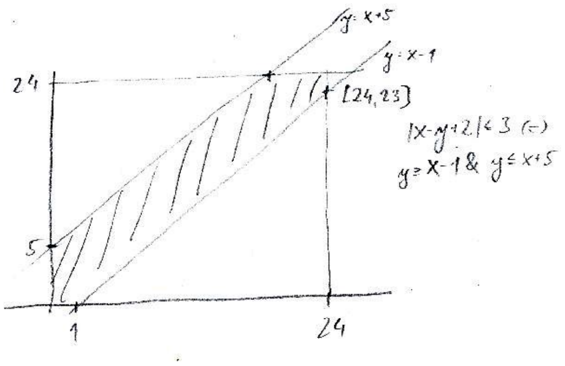


$$p = \frac{24^2 - \frac{1}{2}(23^2 + 19^2)}{24^2} = \frac{131}{576} \doteq 0{,}227.$$

□

1.169. 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)?

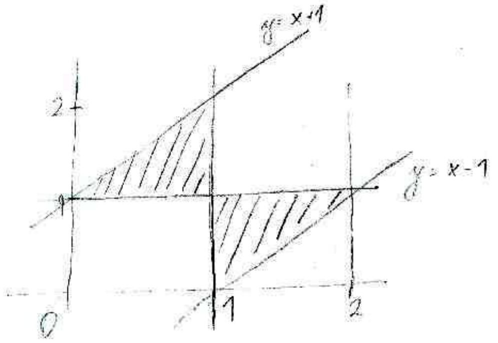
Solution. We are solving analogously to the previous exercise. The space of all events is a square 10×10, Mirek, departing at the time x, meets Marek, departing at the time y if and only if $|x - y| \leq 2$. The probability is $p = \frac{36}{100} = \frac{9}{25} = 0{,}36$.



□

1.170. 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 $1/4$.



□

1.171. Does the equation

$$\text{(a)}\quad \begin{array}{rcrcr} 4x_1 & - & \sqrt{3}x_2 & = & 3, \\ x_1 & - & 2\sqrt{7}x_2 & = & -2; \end{array}$$

$$\text{(b)}\quad \begin{array}{rcrcr} 4x_1 & - & \sqrt{3}x_2 & = & 16, \\ x_1 & - & 2\sqrt{7}x_2 & = & -7; \end{array}$$

$$\text{(c)}\quad \begin{array}{rcrcr} 4x_1 & + & 2x_2 & = & 7, \\ -2x_1 & - & x_2 & = & -3 \end{array}$$

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

$$\begin{vmatrix} 4 & -\sqrt{3} \\ 1 & -2\sqrt{7} \end{vmatrix} = 4 \cdot \left(-2\sqrt{7}\right) - \left(-\sqrt{3} \cdot 1\right) \neq 0,$$

$$\begin{vmatrix} 4 & 2 \\ -2 & -1 \end{vmatrix} = 4 \cdot (-1) - (2 \cdot (-2)) = 0,$$

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. □

1.172. 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

$$S = S_1 + S_2 = \frac{1}{2}\begin{vmatrix} 1-0 & 1-0 \\ -1+2 & 5+2 \end{vmatrix} + \frac{1}{2}\begin{vmatrix} 1-0 & -1-0 \\ 5+2 & 1+2 \end{vmatrix} = \frac{1}{2}(7-1) + \frac{1}{2}(3+7) = 8.$$

□

1.173. 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 the determinants, see 1.34,

$$S = \left|\frac{1}{2}\begin{vmatrix} 1 & 5 \\ 10 & 13 \end{vmatrix}\right| + \left|\frac{1}{2}\begin{vmatrix} 1 & 5 \\ -3 & -5 \end{vmatrix}\right| = \frac{47}{2}.$$

□

1.174. 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

$$\begin{vmatrix} 6-5 & 6-5 \\ 8-5 & 9-5 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 3 & 4 \end{vmatrix} = 1 \cdot 4 - 1 \cdot 3 = 1.$$

□

1.175. 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, $1 + 4 \cdot 1 + \binom{4}{2} \cdot 2 + \binom{4}{3} \cdot 5 + 15 = 52$. □

1.176. Determine the number of ordering relations over a three-element set. ○

1.177. 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). ○

1.178. 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. □

1.179. 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

$$S \circ R = \{(2, 1), (2, 4), (4, 1), (4, 4)\}.$$

□

1.180. 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,

$$R \circ R^{-1} = \{(4, 4), (0, 0), (\pi, \pi), (2, 2), (4, 2), (\pi, 2), (2, \pi), (2, 4)\}.$$

□

1.181. 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

$$(4,2),(2,1)\in R,\quad (4,1)\notin R.$$

□

1.182. 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 $(2,1)$. □

1.183. 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. □

1.184. 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 $(2,-2)\in R$, $(-2,2)\in R$). □

1.185. 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, symmetrical 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.

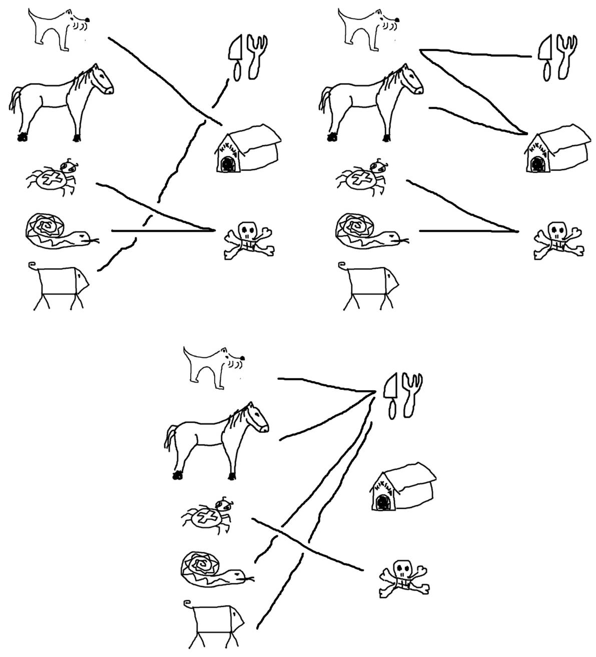
If we chose $X=\{1,2\}$ and the ordering relation

$$R_1=\{(1,1),(2,2),(1,2)\},\quad R_2=\{(1,1),(2,2),(2,1)\}$$

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. □



1.186. 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

$$\{1, 10, 11, \ldots, 18, 19\}, \{2, 20\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}.$$

□

1.187. 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

$$(a, a), (b, b), (c, c), (d, d), (e, e),$$
$$(b, c), (c, b), (a, d), (a, e), (d, a), (d, e), (e, a), (e, d).$$

□

1.188. 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 it is a mapping, and whether it is injective, surjective or bijective.

Solution. In the first case it is a mapping which is surjective but not injective, because both the snake and the spider are labelled as poisonous. The second case is not a mapping but only a relation, since the dog is labelled both as a pet and as a meal. The third case is again a mapping. This time it is neither injective nor surjective. □

1.189. 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). □

1.190. Consider the set of numbers that have five digits in the binary notation and a relation such that two numbers are in the relation 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 $\{17, 18, 20, 24, 23, 27, 29, 30\}$. □

1.191. 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

$$[100] = \{100, 101, 102\} \text{ corresponds } \{9, 10, 11\}$$
$$[110] = \{110, 111, 112\} \text{ corresponds } \{12, 13, 14\}$$
$$[120] = \{120, 121, 122\} \text{ corresponds } \{15, 16, 17\}$$
$$[200] = \{200, 201, 202\} \text{ corresponds } \{18, 19, 20\}$$
$$[210] = \{210, 211, 212\} \text{ corresponds } \{21, 22, 23\}$$
$$[220] = \{220, 221, 222\} \text{ corresponds } \{24, 25, 26\}$$

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

$$[100] = \{100, 200\} \text{ corresponds } \{9, 18\}$$
$$[101] = \{101, 201\} \text{ corresponds } \{10, 19\}$$
$$[102] = \{102, 202\} \text{ corresponds } \{11, 20\}$$
$$[110] = \{110, 210\} \text{ corresponds } \{12, 21\}$$
$$[111] = \{111, 211\} \text{ corresponds } \{13, 22\}$$
$$[112] = \{112, 212\} \text{ corresponds } \{14, 23\}$$
$$[120] = \{120, 220\} \text{ corresponds } \{15, 24\}$$

$$[121] = \{121, 221\} \text{ corresponds } \{16, 25\}$$
$$[122] = \{122, 222\} \text{ corresponds } \{17, 26\}$$

□

1.192. What is the maximal domain D and codomain H 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) $D = \mathbb{R} \setminus \{-1\}$ and $H = \mathbb{R} \setminus \{0\}$. The inverse function is $x \mapsto \frac{1}{x} - 1$.

□

1.193. 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. □

1.194. 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. □

1.195. Let a relation R be defined over the set $N = \{3, 4, 5, \dots, n, n+1, \dots\}$, 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 in 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. □

Solution to the exercises

1.33. $y_n = 2(\frac{3}{2})^n - 2$.

1.92.

i) $(3,3)$, $(4,4)$, $(5,5)$, $(6,6)$, $(7,7)$, $(3,6)$, check that it is an ordering relation.
ii) again (i,i) for $i = 1, \dots, 7$ and additionally $(3,6)$, $(6,3)$, check that it is an equivalence relation.
iii) (i,i) for $i = 1, \dots, 7$ and also $(3,6)$, $(6,3)$, $(4,6)$, $(6,4)$. Check that it is not an equivalence, since transitivity does not hold.

1.103. Three different Hasse diagrams which satisfy the given condition. In total $5! + 5! + 5!/4 = 270$.

1.114.

a) $1 - 3 - 2i + 4i = -2 + 2i$, $1 \cdot (-3) - 8i^2 + 6i + 4i = 5 + 10i$, $1 + 2i$, $\sqrt{4^2 + (-3)^2} = 5$, $\frac{z_1}{z_2} = \frac{z_1 \bar{z}_2}{|z_2|^2} = 1 \cdot (-3) + 8i^2 + 6i - 4i25 = -\frac{11}{25} + \frac{2}{25}i$.
b) $2 + i$, $2i$ 2, 1, $\frac{2}{i} = -2i$.

1.129.

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 $64 - 7 = 57$.

1.139. 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, $p_n = \frac{n}{3}(n^2 - 3n + 8)$.

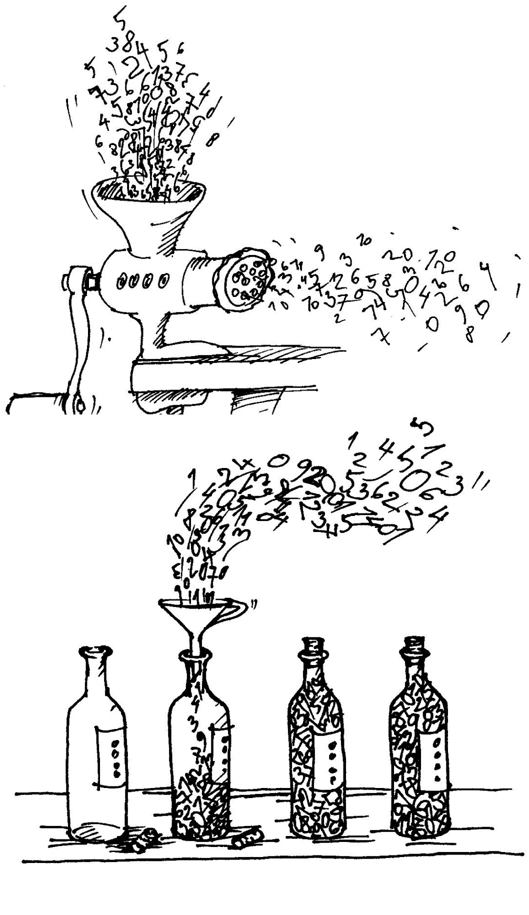
1.176. 19.

1.177. 87.

CHAPTER 2

Elementary linear algebra

are you able to calculate with scalars?
– if not, let us go straight to the matrices...



In the previous chapter we have warmed up with 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 will be dealing with particular topics in more detail.

Three chapters will be about tools for working with that, where the operations consist of simple operations with scalars, but work with more scalars simultaneously. We speak about "linear objects" and "linear algebra". Although it might seem a very specialised tool, we shall see later that even more complicated objects are studied mostly using their "linear approximations".

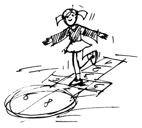
In this chapter we will work directly with finite sequences of scalars. Such sequences arise in real-world problems whenever we have our objects described with more parameters. You should not trouble yourselves with trying to imagine the space with more than three "coordinates". You have to live with the fact that we are able to depict one, two or three dimensions, but we will deal with arbitrary number of dimensions. And when we will observe any parameter in, say, 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 devise tools which will work well even if the number of elements is large.

Also, do not be afraid of terms like field or ring of scalars $\mathbb{K}$. Simply, imagine any specific domain of numbers. Rings of scalars contain for instance integers $\mathbb{Z}$ and all residue classes, among fields we have only $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ and residue classes $\mathbb{Z}_k$ for k prime. Specific among them is $\mathbb{Z}_2$, where from the equation $x = -x$ we cannot infer that $x = 0$, where in every other field we indeed can.

1. Vectors and matrices

Mostly, we speak about vectors in connection with a field of scalars, since the general theory is way more complicated in case there are some non-invertible non-zero scalars. Only in the first two parts of this chapter we will work with vectors and matrices in the context of finite sequences of scalars, and in that case it will be interesting to note how the situation would behave if we had, say, integers instead of a field. It will be hopefully easy to see, how strong results can be derived with precise formal reasoning.

2.1. Vectors over scalars. For now, *vector* is for us an ordered n-tuple of scalars from $\mathbb{K}$, where the fixed $n \in \mathbb{N}$ is called *dimension*.



A. Systems of linear equations

We attack vector spaces from an unexpected direction. We begin with something we know, that is, systems of linear equations. Because even behind them we can actually find vector spaces.

2.1. A colourful example. A company of painters ordered 810 litres of colour, which should contain the same amount of red, green and blue colour (that is, 810 litres of black). The provider can satisfy this order by mixing the colours he usually sells (he has got enough of that in his warehouse), that is

- 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 has to be mixed in order to satisfy the order?

Solution. Denote by

- x – the amount (in litres) of reddish colour to be used;
- y – the amount (in litres) of bluish colour to be used;
- z – the amount (in litres) of greenish colour to be used;

By mixing of 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:

$$0,5x \quad + \quad 0,125y \quad + \quad 0,2z \quad = \quad 270.$$

Analogically, we require (for blue and green colours respectively) that

$$\begin{array}{ccccccc} 0,25x & + & 0,75y & + & 0,2z & = & 270, \\ 0,25x & + & 0,125y & + & 0,6z & = & 270. \end{array}$$

Now we can carry on in two ways. Either we can express individual variables using other variables – from the first equation we have $x = 540{-}0,25y{-}0,4z$, we plug it for x into the second and third equations and obtain two linear equations of two variables $2,75y + 0,4z = 540$ and $0,25y + 2z = 540$. From the second equation we express $z = 270 - 0,125y$ and plugging into the first one we obtain $2,7y = 432$, that is, $y = 160$, therefore $z = 270 - 0,125 \cdot 160 = 250$ and $x = 540 - 0,25 \cdot 160 + 0,4 \cdot 250 = 400$.

Second approach is to use matrix notation. 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

We can add and multiply scalars. We will be able to add vectors, but multiplying a vector will be possible only by scalar. This corresponds to the idea we already saw in the plane $\mathbb{R}^2$, where addition realised vector composition (as of composition of arrows emanating from the origin) and the multiplication by scalar realised stretching of vectors.

Multiple of a vector $u = (a_1, \dots, a_n)$ by a scalar c is defined by multiplying by c every element of the n-tuple u, and also addition is defined coordinate-wise. That means,

Basic vector operations

$$\begin{aligned} u + v &= (a_1, \dots, a_n) + (b_1, \dots, b_n) \\ &= (a_1 + b_1, \dots, a_n + b_n) \\ c \cdot u = c \cdot (a_1, \dots, a_n) &= (c \cdot a_1, \dots, c \cdot a_n). \end{aligned}$$

For vector addition and multiplication by scalars we shall use the same symbols as for scalars, that is 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 (the middle part of the alphabet remains for indices of variables or components and also for summation indices).

We will often require that scalars are from some specific fields, see 1.1, but in this chapter will mostly work with operations that do not require this assumption. In the more advanced literature, this is usually called *modules over rings* instead of vector spaces. In the general theory in the next chapter, we will work exclusively with fields of scalars.

For vector addition in $\mathbb{K}^n$ the properties (KG1)–(KG4) clearly hold with the zero element being

$$0 = (0, \dots, 0) \in \mathbb{K}^n.$$

We are purposely using the same symbol for both the zero vector element and the zero scalar element.

Vector properties

For all vectors $v, w \in \mathbb{K}^n$ and scalars $a, b \in \mathbb{K}$ we have

(V1) $\quad a \cdot (v + w) = a \cdot v + a \cdot w$

(V2) $\quad (a + b) \cdot v = a \cdot v + b \cdot v$

(V3) $\quad a \cdot (b \cdot v) = (a \cdot b) \cdot v$

(V4) $\quad 1 \cdot v = v$

The properties (V1)–(V4) of our vectors, that is, n-tuples of scalars in $\mathbb{K}^n$, are easy to check for any specific ring of scalars $\mathbb{K}$, since when checking the properties we are using for the individual coordinates of vectors only the properties of scalars listed in 1.1 and 1.3.

In this way we shall work with, for instance, $\mathbb{R}^n$, $\mathbb{Q}^n$, $\mathbb{C}^n$, but also with $\mathbb{Z}^n$, $(\mathbb{Z}_k)^n$, $n = 1, 2, 3, \dots$.

2.2. Matrices over scalars. A slightly more complicated object we will use when working with vectors are matrices.

in the third. Therefore the *matrix of the system* is

$$\begin{pmatrix} 0,5 & 0,125 & 0,2 \\ 0,25 & 0,75 & 0,2 \\ 0,25 & 0,125 & 0,6 \end{pmatrix},$$

extended matrix of the system is obtained from the matrix of the system by adding the column of the right-hand sides 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 sequentially doing the so-called elementary row transformations (they correspond to equivalent operations with the equations, see 2.7) we obtain:

$$\left(\begin{array}{ccc|c} 0,5 & 0,125 & 0,2 & 270 \\ 0,25 & 0,75 & 0,2 & 270 \\ 0,25 & 0,125 & 0,6 & 270 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 1 & 3 & 0,8 & 1\,080 \\ 1 & 0,5 & 2,4 & 1\,080 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 2,75 & 0,4 & 540 \\ 0 & 0,25 & 2 & 540 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 11 & 1,6 & 2\,160 \\ 0 & 1 & 8 & 2\,160 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 1 & 8 & 2\,160 \\ 0 & 11 & 1,6 & 2\,160 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0,25 & 0,4 & 540 \\ 0 & 1 & 8 & 2\,160 \\ 0 & 0 & -86,4 & -21\,600 \end{array}\right).$$

And again by plugging the values back we compute

$$\begin{aligned} z &= \frac{-21\,600}{-86,4} = 250, \\ y &= 2\,160 - 8 \cdot 250 = 160, \\ x &= 540 - 0,4 \cdot 250 - 0,25 \cdot 160 = 400. \end{aligned}$$

Thus it is necessary to mix 400 l of reddish, 160 l of bluish and 250 l of greenish colour. □

2.2. Solve a 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 extended matrix of the system

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 2 & -3 & -1 & -3 \\ -3 & 1 & 2 & -3 \end{array}\right),$$

which we using the elementary row transformations transform into the row echelon form

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 2 & -3 & -1 & -3 \\ -3 & 1 & 2 & -3 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -7 & -7 & -7 \\ 0 & 7 & 11 & 3 \end{array}\right) \sim$$

Matrices of type m/n

A matrix of the type m/n over scalars $\mathbb{K}$ is a rectangular schema A with m rows and n columns

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

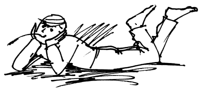
where $a_{ij} \in \mathbb{K}$ for all $1 \le i \le m,\ 1 \le j \le n$. For a matrix A with elements a_{ij} we also use the notation $A = (a_{ij})$.

Vectors $(a_{i1}, a_{i2}, \dots, a_{in}) \in \mathbb{K}^n$ are called (i-th) *rows of the matrix* A, $i = 1, \dots, m$, vectors $(a_{1j}, a_{2j}, \dots, a_{mj}) \in \mathbb{K}^m$ are called (j-th) *columns of the matrix* A, $j = 1, \dots, n$.

Matrix can be also understood as a mapping

$$A : \{1, \dots, m\} \times \{1, \dots, n\} \to \mathbb{K},$$

where $A(i, j) = a_{ij}$. Matrices of the type $1/n$ or $n/1$ are actually just vectors in $\mathbb{K}^n$.



Ever general matrices can be understood as vectors in $\mathbb{K}^{m \cdot n}$, we just concatenate 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 *addition inverse* to the matrix A and the matrix

$$0 = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

is called the *zero matrix*. Seeing matrices as $m \cdot n$-dimensional vectors, we obtain the following claim

Proposition. *The formulas for $A + B$, $a \cdot A$, $-A$, 0 give for the set of all matrices of the type m/n the operations of addition and multiplication by scalars, which satisfy axioms (V1)–(V4).*

2.3. Matrices and equations. An often-used tool for description of mathematical models are systems of linear equations. Matrices are useful for the description of such systems. We use for this the notion of *scalar product* of two vectors, which assigns to the vectors $(a_1, \dots, a_n)$ and $(x_1, \dots, x_n)$ assigns their product

$$(a_1, \dots, a_n) \cdot (x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n$$

that is, we subsequently multiply the coordinates of the vectors and sum the results.

Every system of m linear equations in n variables

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned}$$

can be seen as a constraint on values of m scalar products with an unknown vector $(x_1, \dots, x_n)$ with vectors of coordinates $(a_{i1}, \dots, a_{in})$.

$$\sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 4 & -4 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right).$$

First we have subtracted from the second row twice the first row and to the third row we have added thrice the first row. Then we have added the second row to the third row and multiplied the second row by $-1/4$. Let us now go back to 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 immediately see that $x_3 = -1$. If we plug in $x_3 = -1$ into the equation $x_2 + x_3 = 1$, we obtain $x_2 = 2$. And again by plugging $x_3 = -1$, $x_2 = 2$ into the first equation we obtain $x_1 = 1$. □

Systems of linear equations can be written in the matrix notation. But is it an advantage, when we can solve the systems even without speaking of the matrices? Yes it is, we can speak about the solution with more concept, we can say in the language of matrices how many solutions a system has and it is more natural and elegant for computer applications. Try to get more familiar with particular 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 (this process is called Gaussian elimination, see 2.7), solving the system is then very easy. We show it on some more examples, where we will see that a system can have infinitely many solutions.

2.3. 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 shall 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, which 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:

$$\begin{pmatrix} 2 & -1 & 3 \\ 3 & 16 & 7 \\ 3 & -5 & 4 \\ -7 & 7 & -10 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix}.$$

The vector of variables can be also seen as a column in a matrix of the type $n/1$, and similarly the values $b_1, \dots, b_n$ 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 be then formally written as $A \cdot x = u$ as follows:

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

where the left-hand side is interpreted as m scalar products of the individual rows of the matrix (giving rise to a column vector), whose values are determined by the equations. That means that the identity of the i-th coordinates corresponds to the original i-th equation

$$a_{i1}x_1 + \dots + a_{in}x_n = b_i$$

and the notation $A \cdot x = u$ gives the original system of equations.

2.4. Matrix product. In the plane, that is, for vectors of dimension two, we have developed a matrix calculus and we saw that it is effective to work with (see 1.26). Now we will work more generally and develop all tools we already know from the plane case for all dimensions n.

Matrix multiplication is possible to define only when the dimensions of rows and columns allow it, that is, when the scalar product is defined for them as before:

Matrix product

For any matrix $A = (a_{ij})$ of the type m/n and any matrix $B = (b_{jk})$ of the type n/q over the ring of scalars $\mathbb{K}$ we define their product $C = A \cdot B = (c_{ik})$ as a matrix of the type m/q with the elements

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}, \text{ for arbitrary } 1 \le i \le m, 1 \le k \le q.$$

That is, the element ac[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

$$\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}.$$

2.5. Square matrices. If there is the same number of rows and columns in the matrix, we speak of *square matrix*. The number of rows and columns is then called the *dimension of the matrix*. The matrix

$$E = (\delta_{ij}) = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$$

is called the *unit matrix*. Numbers defined in such way δ_{ij} are also called *Kronecker delta*. Over the set of square matrices over $\mathbb{K}$ of dimension n the matrix product is defined for any two matrices, that is, there is the multiplication operation is defined there, and its properties are similar to that of scalars:

>From there we can see that the second, third and fourth equations are multiples of the equation $7x_2 + x_3 = 0$. But we still carry on with the transformations, in order to see what will happen:

$$\begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 7 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

Although we were given four equations for three variables, the whole system has infinitely many solutions, because for any $x_3 \in \mathbb{R}$ the remaining equations have

$$\begin{aligned} 2x_1 \; - \; x_2 \; + \; 3x_3 &= 0, \\ 7x_2 \; + \; x_3 &= 0 \end{aligned}$$

solution. Thus 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

$$(x_1, x_2, x_3) = (11s, s, -7s), \quad s \in \mathbb{R}.$$

□

2.4. Find all solutions of the system of linear equations

$$\begin{aligned} 3x_1 \qquad\quad + 3x_3 - 5x_4 &= -8, \\ x_1 - x_2 + x_3 - x_4 &= -2, \\ -2x_1 - x_2 + 4x_3 - 2x_4 &= 0, \\ 2x_1 + x_2 - x_3 - x_4 &= -3. \end{aligned}$$

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:

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & -8 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & -3 & 6 & -4 & -4 \\ 0 & 3 & 0 & -2 & -2 \end{array}\right) \sim$$

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 3 & -3 & -3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The system has thus infinitely many solutions, because we have three equations for four variables, which have exactly one solution for any

Proposition. *Over the set of all square matrices of the dimension n over arbitrary ring of scalars $\mathbb{K}$ is defined the multiplication operation with the following properties of rings (see 1.3):*

(1) The associativity holds (O1).
(2) The unit matrix $E = (\delta_{ij})$ is a unit element for multiplication (O3).
(3) The distributivity of multiplication and addition holds (O4).

In general, neither the properties (O2) nor (OI) hold. Therefore, the square matrices for $n > 1$ do not form an integral domain, therefore they are not even a (non-commutative) field.

Proof. Associativity of multiplication – (O1): Since scalars are associative, distributive and commutative, we can for the three matrices $A = (a_{ij})$ of type m/n, $B = (b_{jk})$ of type n/p and $C = (c_{kl})$ of type p/q calculate

$$A \cdot B = \Big(\sum_j a_{ij}.b_{jk}\Big), \quad B \cdot C = \Big(\sum_k b_{jk}.c_{kl}\Big),$$

$$(A \cdot B) \cdot C = \Big(\sum_k \big(\sum_j a_{ij}.b_{jk}\big).c_{kl}\Big) = \Big(\sum_{j,k} a_{ij}.b_{jk}.c_{kl}\Big),$$

$$A \cdot (B \cdot C) = \Big(\sum_j a_{ij}.\big(\sum_k b_{jk}.c_{kl}\big)\Big) = \Big(\sum_{j,k} a_{ij}.b_{jk}.c_{kl}\Big).$$

Note that while computing, we did rely on the fact that it does not matter in which order are we doing given sums and products, that is, we were heavily using the properties of scalars.

We can easily see that multiplication by unit matrix has the property of a unit element:

$$A \cdot E = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & 1 \end{pmatrix} = A$$

similarly for multiplication by E from the left.

It remains to show the distributivity of multiplication and addition. Again using the distributivity of scalars we can easily calculate for matrices $A = (a_{ij})$ of the type m/n, $B = (b_{jk})$ of the type n/p, $C = (c_{jk})$ of the type n/p, $D = (d_{kl})$ of the type p/q

$$\begin{aligned} A \cdot (B + C) &= \Big(\sum_j a_{ij}(b_{jk} + c_{jk})\Big) \\ &= \Big(\big(\sum_j a_{ij}b_{jk}\big) + \big(\sum_j a_{ij}c_{jk}\big)\Big) = A \cdot B + A \cdot C \\ (B + C) \cdot D &= \Big(\sum_k (b_{jk} + c_{jk})d_{kl}\Big) \\ &= \Big(\big(\sum_k b_{jk}d_{kl}\big) + \big(\sum_k c_{jk}d_{kl}\big)\Big) = B \cdot D + C \cdot D. \end{aligned}$$

As we have seen in 1.26, two matrices of dimension two do not necessarily commute:

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

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

choice for the variable $x_4 \in \mathbb{R}$. Thus for x_4 we substitute the parameter $t \in \mathbb{R}$ and go back from the matrix notation to the 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 $x_3 = t - 1$. Plugging in for x_3 into the second equation gives us

$$3x_2 - 3t + 3 + t = 1, \quad \text{that is,} \quad x_2 = \frac{1}{3}(2t - 2)\,.$$

Finally using the first equation we have

$$x_1 - \frac{1}{3}(2t - 2) + t - 1 - t = -2, \quad \text{tj.} \quad x_1 = \frac{1}{3}(2t - 5)\,.$$

Thus the set of solutions can be written (for $t = 3s$) in the form

$$\left\{ (x_1,\, x_2,\, x_3,\, x_4) = \left(2s - \frac{5}{3},\, 2s - \frac{2}{3},\, 3s - 1,\, 3s\right),\ s \in \mathbb{R}\right\}.$$

Let us now go back 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). □

2.5. Determine the solutions of the system of equations

$$\begin{array}{rcrcrcrcr} 3x_1 & & & + & 3x_3 & - & 5x_4 & = & 8, \\ x_1 & - & x_2 & + & x_3 & - & x_4 & = & -2, \\ -2x_1 & - & x_2 & + & 4x_3 & - & 2x_4 & = & 0, \\ 2x_1 & + & x_2 & - & x_3 & - & x_4 & = & -3. \end{array}$$

This gives us immediately a counterexample to validity of (O2) and (OI). For matrices of type 1/1 both axiom clearly holds, because the scalars itself have them. For matrices of greater dimension the counterexamples can be obtained similarly – they have the counterexamples for dimension 2 in their left upper corner, the rest is zero. (Verify it on your own!) □

In the proof we have actually worked with matrices of more general type, thus we have proved the properties in greater generality:

Associativity and distributivity of matrix multiplication

Corollary. *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. Unit matrix is a unit element for multiplication (both from the right and from the left).

2.6. Inverse matrices. With scalars we can do the following: from the equation $a \cdot x = b$ we can express $x = a^{-1} \cdot b$, whenever the inverse of a exists. We would like to be able to do this for matrices too, but we have a problem – how can we tell when such matrix exists, and how to compute it?

We say that B is *inverse* of A, when

$$A \cdot B = B \cdot A = E.$$

Then we write $B = A^{-1}$ and 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 *invertible matrix* or *regular square matrix*.

In the subsequent paragraphs we derive (among other things) that B is inverse of A whenever just one of the required equations holds (the other equation is then necessarily true also).

If A^{-1} and B^{-1} exist, then there also is the inverse of the product $A \cdot B$

$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}. \tag{2.1}$$

Because the associativity of matrix multiplication proved a while ago, we have that

$$(B^{-1} \cdot A^{-1}) \cdot (A \cdot B) = B^{-1} \cdot (A^{-1} \cdot A) \cdot B = E$$
$$(A \cdot B) \cdot (B^{-1} \cdot A^{-1}) = A \cdot (B \cdot B^{-1}) \cdot A^{-1} = E.$$

Because we can calculate with matrices similarly as with scalars (they are just a little more complicated), the existence of 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

$$A \cdot x = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = u$$

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.

$$\left(\begin{array}{cccc|c} 3 & 0 & 3 & -5 & 8 \\ 1 & -1 & 1 & -1 & -2 \\ -2 & -1 & 4 & -2 & 0 \\ 2 & 1 & -1 & -1 & -3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & 8 \end{array}\right) \dots$$

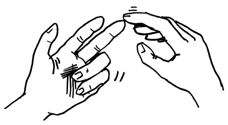
$$\sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 3 & -3 & 13 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 16 \end{array}\right),$$

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 emphasise 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 127

B. Manipulations with matrices

In this sub-chapter we shall work with matrices only, in order to get more familiar with their properties.



2.6. Matrix multiplication. Carry on the matrix multiplications and check the result. 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) $\begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 2 & 1 \\ 1 & 1 & -2 & -3 \\ 3 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 12 & 7 & 1 & -5 \\ 3 & 0 & 5 & 4 \end{pmatrix}$,

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) $\begin{pmatrix} 1 & 3 & -3 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 & 3 \\ -2 & 2 & -1 \\ 3 & 1 & -4 \end{pmatrix} = \begin{pmatrix} -14 & 1 & 12 \end{pmatrix}$,

and when there exists inverse of the matrix A, then we can multiply from the left by A^{-1} and we 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.

2.7. 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 direct solution to the system of equations. But it is important 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.

Let us 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, they become zero), it will also give us some direct informations about the solution. Our simple operations are:

Elementary row transformations

- switching of two rows,
- multiplication of a given row by a non-zero scalar,
- adding a row to another 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 is an integral domain.

Analogically, *elementary column transformations* of matrices are

- switching of two columns
- multiplication of a given column by a non-zero scalar,
- adding a column to another column.

these do not preserve the solution set, since they interchange the variables itself.

Systematically we can use elementary row transformations for subsequent elimination of variables. This gives an algorithm which is usually called *Gaussian elimination* of variables.

Gaussian elimination of variables

Proposition. *Non-zero matrix over arbitrary ring of scalars* $\mathbb{K}$ *can be transformed using finitely many elementary row transformations into the so-called (row) echelon form:*

- *If* $a_{ik} = 0$ *for all* $k = 1, \dots, j$, *then* $a_{kj} = 0$ *for all* $k \geq i$,

vi) $\begin{pmatrix}1 & 2 & -2\end{pmatrix}\cdot\begin{pmatrix}2\\1\\3\end{pmatrix} = \begin{pmatrix}-2\end{pmatrix}.$

Remark. The parts i) and ii) in the previous exercise show that multiplication of square matrices is not commutative in general, in the 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) you can note that $(A \cdot B)^T = A^T \cdot B^T$.

2.7. Calculate A^5 and A^{-3}, if

$$A = \begin{pmatrix}2 & -1 & 1\\ -1 & 2 & -1\\ 0 & 0 & 1\end{pmatrix}.$$

○

2.8. 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 (then we say that such matrices are row equivalent)?

Solution. Both matrices are clearly row equivalent with 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. □

2.9. Find some 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 gradually from the left by elementary matrices (consider what elementary row transformations does it correspond to)

$$E_3 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ -3 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}, \quad E_4 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ -7 & 0 & 0 & 1\end{pmatrix},$$

$$E_5 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1/3 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}, \quad E_6 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & -2 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix},$$

$$E_7 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & -4 & 0 & 1\end{pmatrix}, \quad E_8 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1/4 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix},$$

- *if $a_{(i-1)j}$ is the first non-zero element at the $(i-1)$-th row, then $a_{ij} = 0$.*

Proof. Matrix in the row echelon form looks like this

$$\begin{pmatrix}0 & \dots & 0 & a_{1j} & \dots & \dots & \dots & a_{1m}\\ 0 & \dots & 0 & 0 & \dots & a_{2k} & \dots & a_{2m}\\ \vdots & & & & & & & \\ 0 & \dots & \dots & \dots & \dots & 0 & a_{lp} & \dots\\ \vdots & & & & & & & \end{pmatrix}$$

and the matrix can (but does not have to) end with some zero rows. In order to transform arbitrary matrix we can use a simple algorithm, which will bring us, row by row, to the resulting echelon form:

Gaussian elimination algorithm

(1) By a possible switching of rows we obtain a matrix where the first row has in the first non-zero column a non-zero element, let that column be j-th.
(2) For $i = 2, \dots$, by multiplying the first row by the element a_{ij}, multiplying i-th row by the element a_{1j} and subtracting we eliminate the element a_{ij} 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 proves the proposition. □

The given algorithm is really the usual elimination of variables used in the systems of linear equations.

In a completely analogical manner we define the column echelon form of matrices and doing column instead of row elementary transformations, we obtain an algorithm for transforming matrices into the column echelon form.

Remark. We have formulated the Gaussian elimination for general scalars from some ring. It seems natural to multiply with scalars to obtain a row echelon form where the coefficients at the non-zero "diagonal" are ones – computing the solution is then easy. However, this is not possible in general – take for instance the integers $\mathbb{Z}$.

For solving systems of equations the given algorithm does not make any sense if there are divisors of zero among the scalars. Think carefully about the differences between $\mathbb{K} = \mathbb{Z}$, $\mathbb{K} = \mathbb{R}$ and possibly $\mathbb{Z}_2$ or $\mathbb{Z}_4$.

2.8. Matrix of elementary row transformations. In the following we will work exclusively with field of scalars $\mathbb{K}$, that is, every non-zero scalar has an inverse.

Note that elementary row (column) transformations correspond to multiplication from the left (right) by the following matrices:

we obtain

$$B = E_8E_7E_6E_5E_4E_3E_2E_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1/12 & -5/12 & 0 \\ 1 & -2/3 & 1/3 & 0 \\ 0 & -4/3 & -1/3 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & -3 & -5 & 0 \\ 0 & 1 & 9/4 & 1/4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

□

2.10. Complex numbers as matrices. Consider the set of matrices $C = \{\begin{pmatrix} a & b \\ -b & a \end{pmatrix}, a, b \in \mathbb{R}\}$. 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 then called 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 $(3 + 5i) \cdot (8 - 9i) = 69 - 13i$.

2.11. 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. It has to hold

$$\begin{pmatrix} a_1 + 3c_1 & b_1 + 3d_1 \\ 3a_1 + 8c_1 & 3b_1 + 8d_1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

that is,

$$\begin{array}{rrrrl} a_1 & & + \; 3c_1 & & = 1, \\ & b_1 & & + \; 3d_1 & = 2, \\ 3a_1 & & + \; 8c_1 & & = 3, \\ & 3b_1 & & + \; 8d_1 & = 4. \end{array}$$

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

(1) Switching of the i-th and j-th row (column)

$$\begin{pmatrix} 1 & 0 & \dots & & & & \\ 0 & \ddots & & & & & \\ \vdots & & 0 & \dots & 1 & & \\ & & \vdots & \ddots & \vdots & & \\ & & 1 & \dots & 0 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

.

(2) Multiplication of the i-th row (column) by the scalar a:

$$\begin{array}{cl} \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ \\ \leftarrow i \\ \\ \\ \\ \end{array} \end{array}$$

.

(3) Adding the i-th and the j-th row (column):

$$\begin{array}{cc} \begin{array}{c} \\ \\ \\ \\ i \to \\ \\ \\ \end{array} & \begin{pmatrix} 1 & 0 & & & & & \\ 0 & \ddots & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & 1 & & \ddots & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}. \\ & \begin{array}{c} \uparrow \\ j \end{array} \end{array}$$

This trivial observation is actually very important, since the product of invertible matrices is invertible (recall 2.1) and all elementary transformations over 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 corresponding matrix).

For arbitrary matrix A we obtain by multiplying with suitable invertible matrix $P = P_k \cdots P_1$ from the left (that is, sequential multiplication with k matrices) its equivalent row echelon form $A' = P \cdot A$.

In general, if we apply the same elimination procedure for the columns, we can obtain from any matrix B 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 we will use many times in the future:

2.9. Theorem. *For every matrix A of the type m/n over field of scalars $\mathbb{K}$ there exists square invertible matrices P of dimension m and Q of dimension n such that the matrix $P \cdot A$ is in row echelon*

(-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 find the values a_2, b_2, c_2, d_2 by a different approach. We use the relation

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix},$$

which holds for any numbers $a, b, c, d \in \mathbb{R}$ (easy to derive; it also directly follows from 2.2), 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

$$X_2 = \begin{pmatrix}-2 & 1\\ -12 & 5\end{pmatrix}.$$

□

2.12. Solve the matrix equation

$$X \cdot \begin{pmatrix}2 & 5\\ 1 & 3\end{pmatrix} = \begin{pmatrix}4 & -6\\ 2 & 1\end{pmatrix}.$$

○

2.13. Computing the inverse matrix. Compute the inverse matrix 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: 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 gradually obtain

$$\left(\begin{array}{ccc|ccc}4 & 3 & 2 & 1 & 0 & 0\\ 5 & 6 & 3 & 0 & 1 & 0\\ 3 & 5 & 2 & 0 & 0 & 1\end{array}\right) \sim \left(\begin{array}{ccc|ccc}1 & -2 & 0 & 1 & 0 & -1\\ 5 & 6 & 3 & 0 & 1 & 0\\ 3 & 5 & 2 & 0 & 0 & 1\end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc}1 & -2 & 0 & 1 & 0 & -1\\ 0 & 16 & 3 & -5 & 1 & 5\\ 0 & 11 & 2 & -3 & 0 & 4\end{array}\right) \sim \left(\begin{array}{ccc|ccc}1 & -2 & 0 & 1 & 0 & -1\\ 0 & 5 & 1 & -2 & 1 & 1\\ 0 & 11 & 2 & -3 & 0 & 4\end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc}1 & -2 & 0 & 1 & 0 & -1\\ 0 & 5 & 1 & -2 & 1 & 1\\ 0 & 1 & 0 & 1 & -2 & 2\end{array}\right) \sim \left(\begin{array}{ccc|ccc}1 & 0 & 0 & 3 & -4 & 3\\ 0 & 0 & 1 & -7 & 11 & -9\\ 0 & 1 & 0 & 1 & -2 & 2\end{array}\right)$$

form and

$$P \cdot A \cdot Q = \begin{pmatrix}1 & \dots & 0 & \dots & \dots & \dots & 0\\ \vdots & \ddots & & & & & \\ 0 & \dots & 1 & 0 & \dots & \dots & 0\\ 0 & \dots & 0 & 1 & 0 & \dots & 0\\ 0 & \dots & 0 & 0 & 0 & \dots & 0\\ \vdots & & & & & & \end{pmatrix}.$$

2.10. Algorithm for computing inverse matrix. In the previous paragraphs we have basically obtain the complete algorithm for computing the inverse matrix. Using the simple approach described in the next paragraph, we either find out that the inverse does not exist, or we compute it. Keep in mind that we are still working over field of scalars.

Equivalent row transformations of square matrix A of dimension n lead to the matrix P' such that the matrix $P' \cdot A$ is in row echelon form. It could be the case that some of the last rows are zero. If there exists the inverse of A, 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 B of dimension n. That is, the existence of zero row in the result of (row) Gaussian elimination means that there cannot exists the inverse A^{-1}.

Assume now that A^{-1} exists. Because of the previous, we obtain the row echelon form which has no non-zero row, that is, all diagonal elements of $P' \cdot A$ are non-zero. But then carrying the row elimination using the elementary row transformation from the bottom-right corner backwards and transforming the diagonal elements to be units we obtain the unit matrix E. That is, we find another invertible matrix P'' such that for $P = P'' \cdot P'$ we have $P \cdot A = E$. Doing column instead of row transformation we can (under the assumption of the existence of A^{-1}) find a matrix Q such that $A \cdot Q = E$. From this we have

$$P = P \cdot E = P \cdot (A \cdot Q) = (P \cdot A) \cdot Q = Q.$$

That is, we have found the inverse matrix

$$A^{-1} = P = Q$$

for the matrix A. Notably, at the point of finding the matrix P with the property $P \cdot A = E$ we don't have to do any further computation since we know that we already have the inverse matrix.

Practically we can work as follows:

COMPUTING THE INVERSE MATRIX

Next to each other we write the original matrix A and the unit matrix E, we transform the matrix A using the elementary row transformation to the row echelon form, then using the so-called backwards elimination to the diagonal matrix and then by multiplying with the inverse elements of $\mathbb{K}$ to the unit matrix. Simultaneously, we apply all these transformations to the matrix E, and as a result we obtain the matrix A^{-1} in place where E has been. If during the course of the procedure we obtain a zero row in the original matrix, we conclude that the inverse matrix does not exist.

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3 & -4 & 3 \\ 0 & 1 & 0 & 1 & -2 & 2 \\ 0 & 0 & 1 & -7 & 11 & -9 \end{array}\right).$$

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. Let us emphasise the result

$$A^{-1} = \begin{pmatrix} 3 & -4 & 3 \\ 1 & -2 & 2 \\ -7 & 11 & -9 \end{pmatrix}.$$

Let us 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

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 3 & 3 & 4 & 0 & 1 & 0 \\ 2 & 2 & 3 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 3 & 1 & -3 & 1 & 0 \\ 0 & 2 & 1 & -2 & 0 & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 3 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1/3 & 0 & -2/3 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & \frac{1}{3} & -1 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & -\frac{2}{3} & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 2 & -3 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & \frac{1}{3} & 0 & -\frac{2}{3} & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 2 & -3 \\ 0 & 1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 0 & -2 & 3 \end{array}\right),$$

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

$$\left(A^T \cdot B\right)^{-1} = \begin{pmatrix} 1 & 2 & -3 \\ -1 & 1 & -1 \\ 0 & -2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 & -7 \\ -4 & -2 & 11 \\ 3 & 2 & -9 \end{pmatrix}$$

$$= \begin{pmatrix} -14 & -9 & 42 \\ -10 & -5 & 27 \\ 17 & 10 & -49 \end{pmatrix}.$$

□

2.11. Linear dependence and rank. In the previous musings about calculations with matrices we have worked all the time with row and column addition seeing them as vector, along with scalar multiplication. Such operations are called *linear combinations*. We return to such operations in abstract sense in a while in 2.24, it will be also useful to understand their core meaning just now. Linear combination of rows (columns) of a matrix $A = (a_{ij})$ of type m/n we understand an expression of the form

$$c_1 u_{i_1} + \cdots + c_k u_{i_k},$$

where c_i are scalars, $u_j = (a_{j1}, \dots, a_{jn})$ are rows (or $u_j = (a_{1j}, \dots, a_{mj})$ are columns) of the matrix A.

If there exists linear combination of given rows with at least one non-zero scalar coefficient which results into zero row, we say that these rows are *linearly dependent*. In the other case, that is, when the only possibility of obtaining zero row is by taking only zero scalars, 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 2.9 with h ones on the diagonals and assume that by two different transformation sequences we obtain two different $h' < h$. But then according to our algorithm there are two invertible matrices P and Q such that

$$P \cdot E_{h'} \cdot Q = E_h.$$

In the product $E_{h'} \cdot Q$ there will be more zero rows in the bottom part of the matrix than ones in E_h – but we should be able to reach E_h using only elementary row transformations. Increasing the number of linearly independent rows using only elementary row transformations is not possible. Therefore the number of ones in the matrix $P \cdot A \cdot Q$ in the theorem 2.9 is independent of the choice of our elimination sequence and is always equal to the number of linearly independent rows in A, and also to the number of linearly independent columns in A. This number is called the *rank of the matrix* and we denote it by $h(A)$. Let us remember the following theorem:

Theorem. *Let A be a matrix of type m/n over field of scalars $\mathbb{K}$. The matrix A has the same number $h(A)$ of linearly independent rows and columns. Notably, the rank is always at most minimum of the dimensions of the matrix A.*

The algorithm for computing the inverse matrix also says that square matrix A of dimension m has inverse if and only if its rank equals m.

2.12. Matrices as mappings. Analogous to the way we did it when working with matrices in the geometry of the plane (see 1.29), we can interpret every square matrix A as a mapping

$$A : \mathbb{K}^n \to \mathbb{K}^n, \quad x \mapsto A \cdot x.$$

Thanks to the distributivity of matrix multiplication it is clear how are linear combinations of vectors mapped using such mappings:

$$A \cdot (a\,x + b\,y) = a\,(A \cdot x) + b\,(A \cdot y).$$

2.14. Compute the inverse matrix of the matrix

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 2 & -2 & 1 \\ 5 & -5 & 2 \end{pmatrix}.$$

2.15. Compute the inverse matrix of the matrix

$$\begin{pmatrix} 8 & 3 & 0 & 0 & 0 \\ 5 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 & 5 \end{pmatrix}.$$

2.16. 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}.

2.17. Compute A^{-1}, if

(a) $A = \begin{pmatrix} 1 & i \\ -i & 3 \end{pmatrix}$, while i is the imaginary unit

(b) $A = \begin{pmatrix} 1 & -5 & -3 \\ -1 & 5 & 4 \\ -1 & 6 & 2 \end{pmatrix}$.

2.18. Write the inverse matrix to the $n \times n$ matrix ($n > 1$)

$$A = \begin{pmatrix} 2-n & 1 & \cdots & 1 & 1 \\ 1 & 2-n & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & \ddots & \ddots & 2-n & 1 \\ 1 & 1 & \cdots & 1 & 2-n \end{pmatrix}.$$

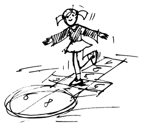
C. Permutations

In order to be able to define the key point of the matrix calculus, that is, determinant, we must deal with permutations (bijections of a finite set) and their parities.

We shall use the two-row notation for permutations (see 2.14). In the first row we list all elements of the given set, and every column

Right from the definition we see (thanks to the associativity of multiplication) that composition of mappings corresponds to matrix multiplication in given order. Thus invertible matrices correspond to bijective mappings.

>From this point of view the theorem 2.9 is very interesting.



We can see it as follows: the rank of the matrix determines how big is the image of the whole $\mathbb{K}^n$ under this mapping. Really, if $A = P \cdot E_k \cdot Q$ with matrix E_k with k ones as in 2.9, then the invertible Q first just bijectively "shuffles" the n-dimensional vectors in $\mathbb{K}^n$, the matrix E_k then "copies" first k coordinates and zeroes the $n-k$ remaining. This "k-dimensional" image then cannot be enlarged by multiplying with P.

2.13. Solving systems of linear equations. We shall return to the

notions of dimension, linear independence and so on in the third part of this chapter. But even now we can notice what the already derived results say about the solution of the system 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 approach we have presented before corresponds to sequential variable elimination in the equations and deletion of the linearly dependent equations (these are simply a consequence of other equations).

We have thus derived the 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 nonzero rows in the row echelon form of the extended matrix than in original matrix of the system, then there cannot be any solution (simply, we cannot hit the given value with the corresponding linear mapping). If the rank of both matrices is the same, then we have in the backwards elimination exactly that many free parameters as the difference between the number of variables n and the rank $h(A)$.

2. Determinants

In the fifth part of the first chapter we have seen (see 1.27) that for square matrices of dimension 2 over the real numbers there exists scalar function det, which assigns to the matrix a non-zero number if and only if the inverse of the matrix exists. We did not say it in these words, but you can check by yourself that it means indeed the same (see the paragraphs starting with 1.26 and the formula (1.16)). Determinant was also useful in another way, see the paragraphs 1.33 and 1.34, where we have derived that the volume of the parallelepiped should be linearly dependent on every of the two vectors defining it and it is useful to require the change of the sign when changing the order of these vectors. Because determinant (and only determinant) had these properties, up to a constant scalar multiple, we stated that it corresponds to the definition of the volume. Now we will see that we can do it similarly for every finite dimension.

In this part we will work 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.

2.14. Definition of the determinant. Let us remind that the bijective mapping from the set X to itself is called *permutation of* then corresponds to a tuple (preimage, image) in the given permutation. Because permutation is a bijection, the second row is indeed a permutation (ordering) of the first row, in accordance with the definition from combinatorics.

2.19. 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: let us start with the first element (one) and look on the second row to see what the image of one is. It is three. Now we look on the column that starts with three, and find out that the image of three is six, and so on. We carry on in this manner for so long until we again reach the starting element (in this case it is one). 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 cycle (see 2.16) which we denote $(1, 3, 6, 9, 2)$.

Now we take any element not contained in the obtained cycle, and start with him the same procedure as with one. 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 ($\{1, 2, \dots, 9\}$) appears in one of the obtained cycles, we can thus write:

$$\sigma = (1, 3, 6, 9, 2) \circ (4, 7, 5, 8).$$

For cycles the decomposition into transpositions is simple, we have

$$(1, 3, 6, 9, 2) = (1, 3)\circ(3, 6)\circ(6, 9)\circ(9, 2) = (1, 3)(3, 6)(6, 9)(9, 2).$$

We thus obtain:

$$\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8).$$

□

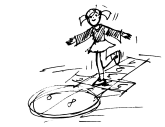
Remark. Let us note that the operation $\circ$ is composition of mappings, thus it is necessary to carry out the composition "backwards", as we are used to composition of mappings. Applying the given composition of transposition for instance on the element two we can gradually write:

$$\begin{aligned}[(1, 3)(3, 6)(6, 9)(9, 2)](2) &= [(1, 3)(3, 6)(6, 9)]((9, 2)(2)) = \\ &[(1, 3)(3, 6)(6, 9)](9) \\ &= [(1, 3)(3, 6)](6) = (1, 3)(3) = 1,\end{aligned}$$

the set X, see 1.7. If $X = \{1, 2, \dots, n\}$, the permutation can be written putting the resulting ordering into a table:

$$\begin{pmatrix} 1 & 2 & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(n) \end{pmatrix}.$$

The element $x \in X$ is called a fixed point of the permutation σ if $\sigma(x) = x$. Permutation σ such that 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, is called *transposition*, we denote it by (x, y). Of course that for such transformation it holds also that $\sigma(y) = x$, therefore the name.



In the dimension 2 the formula for 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 switching two columns leads to the change of the sign of the whole result, and sum all of them (that is, all two):

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

In general, consider 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:

Definition of determinant

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 $\{1, \dots, n\}$ and the sign sgn for a permutation σ will be described later. Each of the expressions

$$\operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

is called a *member of the determinant* $|A|$.

In the dimensions 2 and 3 we can easily guess correct signs. The product of the elements on the diagonal should be with positive sign and we want anti-symmetry when switching two columns or rows.

Determinants in the dimension 2 and 3

For $n = 2$ it is, as we have expected

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Similarly for $n = 3$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{matrix} a_{11}a_{22}a_{33} - a_{13}a_{22}a_{31} + a_{13}a_{21}a_{32} \\ -a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33}. \end{matrix}$$

This formula is called the *Saarus rule.*

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 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.

Let us further note 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.

2.20. Determine the parity of the following permutations:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 1 & 6 & 7 & 8 & 9 & 5 & 4 & 2 \end{pmatrix} \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 1 & 5 & 3 \end{pmatrix}$$

Solution. >From the previous exercise we know that $\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8)$. Its parity is given by the parity of the number of transpositions in its decomposition (which is, unlike the number of transposition in an arbitrary decomposition, always the same). There are seven transpositions in the decomposition, thus the permutation is odd. Even without the knowledge of a decomposition of σ into transpositions we could compute the number of tuples $(a, b) \subset \{1, 2, \dots, 9\} \times \{1, 2, \dots, 9\}$ which are inverse with respect to σ (see 2.15): 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 realise that the number of inversions in a given permutation is exactly the number of tuples "bigger 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 4, 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 the permutation is indeed odd.

2.15. Parity of permutation. How can we find a sign of a permutation. We say that a tuple of element $a, b \in X = \{1, \dots, n\}$ forms an *inversion in permutation*σ, if $a < b$ and $\sigma(a) > \sigma(b)$. Permutation σ is called *even* (*odd*), if it contains even (odd) number of inversions.

Parity of 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 would like to know how to calculate with parity. >From the following theorem about permutations it is clear that the Saarus rule really gives the determinant for the dimension 3.

Theorem. *Over the set $X = \{1, 2, \dots, n\}$ there are exactly $n!$ distinct permutations. These can be ordered in a sequences such that every two consequent permutations differ in exactly one transposition. For any permutation there is such sequence starting with it. Every transposition changes parity.*

Proof. For one- and two-element X the claim is trivial. Let us do induction over the number of dimensions.

Assume that the claim holds for all sets with $n-1$ elements and consider a permutation $\sigma(1) = a_1, \dots, \sigma(n) = a_n$. According to the induction assumption all the permutations that end with a_n can be obtained in a sequence where every two consequent permutations differ in one transposition. There are $(n-1)!$ such permutations. On the last of them we use the transposition of $\sigma(n) = a_n$ with some element a_i has not yet been at the last position, and once again form a sequence of all permutations that end with a_i. After doing this procedure n-times, we obtain $n(n-1)! = n!$ distinct permutations – that is, all permutations on n elements. The resulting sequence satisfies the condition.

Note that the last sentence of the theorem does not seem to be useful for its application. But it is a very important part for proving it by induction over the size of X.

It remains to prove the part of the theorem about parities. Consider the ordering

$$(a_1, \dots, a_i, a_{i+1}, \dots, a_n),$$

containing r inversions. Then clearly in the ordering

$$(a_1, \dots, a_{i+1}, a_i, \dots, a_n)$$

there are either $r-1$ or $r+1$ inversions. Every transposition (a_i, a_j) is obtainable by doing $(j-i) + (j-i-1) = 2(j-i) - 1$ transpositions of neighbouring elements. Therefore doing any transposition changes the parity. Also, we already know that all permutations can be obtained by applying transpositions. □

We found out that applying a transposition changes the parity of a permutation and any ordering of numbers $\{1, 2, \dots, n\}$ can be obtained through transposing of neighbouring elements. Therefore we have proven

Corollary. *On every finite set $X = \{1, \dots, n\}$ with n elements, $n > 1$, there are exactly $\frac{1}{2}n!$ even and $\frac{1}{2}n!$ odd permutations.*

If we compose two permutations, it means first doing all transpositions forming the first permutations and then all the transpositions forming the second. Therefore for any two permutations $\sigma, \eta : X \to X$ we have

$$\operatorname{sgn}(\sigma \circ \eta) = \operatorname{sgn}(\sigma) \cdot \operatorname{sgn}(\eta)$$

Analogously we can decompose τ into either a product of 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: $1 + 2 + 3 + 0 + 1 = 7$. Anyway we find out that τ is also an odd permutation.

□

D. Determinants

Ensure on the following exercise that you can compute determinants of the type 2×2 and 3×3 (using the Saarus rule):

2.21. Compute the determinant of the following matrices $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 2 & 2 \end{pmatrix}$ $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix}$.

2.22. 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 start by expanding the first column, where there is greatest number (one) of zeroes. Gradually we get

$$\begin{vmatrix} 1 & 3 & 5 & 6 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{vmatrix} = 1 \cdot \begin{vmatrix} 2 & 2 & 2 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} - 1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} + 1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 2 & 2 & 2 \\ 1 & 2 & 1 \end{vmatrix}$$

$$\overset{\text{using the Saarus rule}}{=} -2 - 2 + 6 = 2.$$

□

2.23. Find all the values of argument 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:

$$D = \begin{vmatrix} a & 1 & 1 & 1 \\ 0 & a & 1 & 1 \\ 0 & 1 & a & 1 \\ 0 & 0 & 0 & -a \end{vmatrix} = a \cdot \begin{vmatrix} a & 1 & 1 \\ 1 & a & 1 \\ 0 & 0 & -a \end{vmatrix},$$

further we expand using the last row:

$$D = a \cdot (-a) \begin{vmatrix} a & 1 \\ 1 & a \end{vmatrix} = -a^2(a^2 - 1).$$

and also

$$\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma).$$

2.16. Decomposing permutations into cycles. A good tool for practical work with permutations is the cycle decomposition.

CYCLES

Permutation σ over the set $X = \{1, \dots, n\}$ is called *cycle* of length k, if we can find elements $a_1, \dots, a_k \in X$, $2 \le k \le n$ such that $\sigma(a_i) = a_{i+1}$, $i = 1, \dots, k-1$, while $\sigma(a_k) = a_1$ and other elements in X are fixed-points of σ. Cycles of length two are exactly transpositions.

Every permutation is a composition of cycles. Cycles of even length have parity -1, cycles of odd length have parity 1.

The last claim has yet to be proven. If we define for a given permutation σ relation r such that two elements x, $y \in X$ are in relation if and only if $\sigma^r(x) = y$ for some iteration of the permutation σ, then clearly it is an equivalence relation (check it carefully!). Because X is finite set, for some ℓ it must hold that $\sigma^\ell(x) = x$. If we pick one equivalence class $\{x, \sigma(x), \dots, \sigma^{\ell-1}(x)\} \subset X$ and define other elements to be fixed-points, we obtain a cycle. Evidently, the original permutation X is then 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 an odd number of transposition, therefore their parity is -1. Analogously, cycle of odd length can be obtained using an even number of transpositions and therefore it has parity 1.

2.17. Simple properties of determinant. Knowing the properties of permutations and their parities from previous paragraphs allows us to derive quickly some basic properties of determinant.

For every matrix $A = (a_{ij})$ of the type m/n over scalars from $\mathbb{K}$ we define *transpose* of A. It is a matrix $A^T = (a'_{ij})$ with elements $a'_{ij} = a_{ji}$ which is of the type n/m.

Square matrix A with the property $A = A^T$ is called *symmetric*. If it holds that $A = -A^T$, then A is called *antisymmetric*.

SIMPLE PROPERTIES OF DETERMINANT

Theorem. *For every square matrix $A = (a_{ij})$ the following claims hold:*

(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 a 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) Determinant $|A|$ does not change if we add to any row of A a linear combination of other rows.

Together we obtain the following condition for a: $a^4 - a^2 + 1 = 0$. Substituting $t = a^2$ we have $t^2 - t + 1$ with roots $t_1 = \frac{1+i\sqrt{3}}{2} = \cos(\pi/3) + i\sin(\pi/3)$, $t_1 = \frac{1-i\sqrt{3}}{2} = \cos(\pi/3) - i\sin(\pi/3) = \cos(-\pi/3) + i\sin(-\pi/3)$, from where we obtain four possible values for the parameter a: $a_1 = \cos(\pi/6) + i\sin(\pi/6) = \sqrt{3}/2 + i/2$, $a_2 = \cos(7\pi/6) + i\sin(7\pi/6) = -\sqrt{3}/2 - i/2$, $a_3 = \cos(-\pi/6) + i\sin(-\pi/6) = \sqrt{3}/2 - i/2$, $a_4 = \cos(5\pi/6) + i\sin(5\pi/6) = -\sqrt{3}/2 + i/2$. □

2.24. Vandermonde determinant. Prove the formula for the so-called Vandermonde determinant, that is, determinant of the Vandermonde matrix:

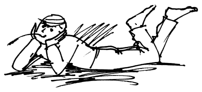
$$V_n = \begin{vmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_n \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & & \vdots \\ a_1^{n-1} & a_2 & \dots & a_n^{n-1} \end{vmatrix} = \prod_{1 \le i < j \le n} (a_j - a_i),$$

where $a_1, \dots, a_n \in \mathbb{R}$ and on the right-hand side of the equation there is the product of all terms $a_j - a_i$ where $j > i$.

Solution.

We show a really beautiful proof by induction, which fills the heart of any mathematician with supreme joy. Consider the determinant V_n to be a polynomial P in the variable a_n. From the definition of the determinant it follows that this polynomial is of degree $n-1$ in this variable and that the numbers $a_1,\dots,a_{n-1}$ are its roots: if we substitute in the Vandermonde matrix V_n into the last column formed by the powers of a_n any of the previous columns formed by the powers of the number a_i, the value of this changed determinant is actually the value of the Vandermonde determinant (seen as the polynomial in the variable a_n) at the point a_i. However, that determinant is clearly zero, because determinant of matrix with two identical, that is, linearly dependent columns is zero. That means that a_i is a root of P. Thus we have $n-1$ roots of a polynomial of degree k, thus it must be the list of all its roots and P must be of the form $P = C(a_n - a_1)(a_n - a_2)\cdots(a_n - a_{n-1})$ where C is some constant – the leading term of the polynomial P. If we consider computation of the determinant V_n using the last column expansion, we see that C is the coefficient at a_n^{n-1}, that is $V(n-1)$. Since for $n = 2$, clearly $V(2) = a_2 - a_1$, the laim for V_n holds by induction. □

Alternative solution. (see Hints and solutions to the exercises)



Proof. (1) The members of determinants $|A|$ and $|A^T$ are in bijective correspondence. To a member $\mathrm{sgn}(\sigma)a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$ corresponds in A^T member (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. The parity of σ and σ^{-1} is the same, therefore it really is a member in the determinant of $|A^T|$ and the first claim is proven.

(2) This comes straight from the definition of determinant, because all its members contain from every row exactly one member. If one of the rows is zero, all members of the determinant are zero.

(3) In all members of $|B|$ the only change in comparison with $|A|$ is an addition of one transposition, therefore all the signs will be reversed.

(4) This is straight from the definition, because members of $|B|$ are members of $|A|$ multiplied by the scalar a.

(5) In every member of $|A|$ there is exactly one element from the k-th row of the matrix A. Thanks to 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, among the members of determinant there are always two identical up to the sign. Therefore in this case $|A| = 0$. Thanks to the claim (5), we can add to the given row any other row without changing the value of the determinant. Thanks to the claim (5), we can add even a scalar multiple of any other row. □

2.18. Corollaries for computation. Thanks to the previous theorem, we can using elementary row transformations bring every square matrix A into row echelon form, without changing the value of its determinant. We just have to be careful and add to rows only linear combinations of other rows.

Computing determinants using elimination

If the matrix A is in row echelon form, then every member of $|A|$ at least one element lies below the diagonal, except for the case that all elements lie on the diagonal. Therefore the diagonal member is the only non-zero one. Thus we see that the determinant of such matrix in row echelon form is

$$|A| = a_{11} \cdot a_{22} \cdot \dots \cdot a_{nn}.$$

Previous theorem gives us very effective method for computing determinants using Gauss elimination method, see the paragraph 2.7.

Let us note a nice corollary of the first claim of the previous theorem about the equality of the determinant of the matrix and its transpose. It ensures that whenever we prove some claim about determinants formulated in terms of rows of the corresponding matrix, we immediately obtain an analogous claim in terms of the columns. For instance, we can immediately formulate all the claims (2)–(6) for

2.25. Find out whether 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. 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 (2.32) 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. □

E. Systems of linear equations for the second time

We have already encountered systems of linear equations at the beginning of the chapter. Now we will deal with them in more detail. Let us first use the advantage for computing the solution of the system of linear equations given by the inverse of the matrix.

2.26. Participants of a trip. There were 45 participants of a two-day bus trip. First day the fee for a watchtower was €30 for an adult, €16 for a child and €24 for a senior. In total, the fee 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 c for a child and €34 for a senior. In total, the fee was €1 542. How many adults, children and seniors were there among the participants?

Solution. Let us introduce the variables

x giving the „number of adults“;

y giving the „number of children“;

z giving the „number of seniors“;

There were 45 participants, therefore

$$x \;+\; y \;+\; z \;=\; 45.$$

addition of linear combinations of columns. We can use it for deriving the following formula for direct calculation of solutions for systems of linear equations:

Cramer rule

Consider the system of n linear equations for n variables with matrix of the system $A = (a_{ij})$ and the column of values $b = (b_1, \dots, b_n)$, that is, in matrix notation we are solving the equation $A \cdot x = b$. If there exists the inverse A^{-1}, then individual components of the unique solution $x = (x_1, \dots, x_n)$ given by the relation

$$x_i = |A_i||A|^{-1},$$

where the matrix A_i arises from the matrix of the system A by exchanging the i-th column for the column b of values.

Really, as we have already seen, inverse of the matrix of the system exist if and only if the system has unique solution. If we have such solution x, we can plug instead of the column b into the matrix A_i the corresponding linear combination of the columns of the matrix A, that is the values $b_i = a_{i1}x_1 + \dots + a_{in}x_n$. Then, by subtracting the x_k-multiples of all other columns, in the i-th column remains 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 $|A_i||A|^{-1} = x_i|A||A|^{-1} = x_i$, which is the claim.

Let us further note that the properties (3)–(5) from the previous theorem say that determinant, as a mapping which assign to n vectors of dimension n (rows or columns of the matrix) a scalar, is antisymmetric mapping linear in every argument, exactly as we required in analogy to the 2-dimensional case.

2.19. 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 over $\mathbb{R}^n$ we can see the determinant of this matrix as expression of the ratio between the volume of the parallelepipeds given by the vectors $x_1, \dots x_n$ and their images $A \cdot x_1, \dots, A \cdot x_n$. Because mapping composition $x \mapsto A \cdot x \mapsto B \cdot (A \cdot x)$ corresponds to matrix multiplication, the so-called *Cauchy theorem* is easy to understand:

Cauchy theorem

Theorem. *Let $A = (a_{ij})$, $B = (b_{ij})$ be square matrices of dimension n over the ring of scalars $\mathbb{K}$. Then $|A \cdot B| = |A| \cdot |B|$.*

Not that from the Cauchy theorem and the representation of the elementary row transformations by multiplication by suitable matrices (see 2.8) we immediately have the claim (2), (3) and (6) from the theorem 2.17.

We know derive this theorem in a purely algebraic way just because the previous argumentation based on geometrical intuition could hardly work for arbitrary scalars. The base tool is the so-called *determinant expansion* using one or more of the rows or columns. We will also need a little of technical preparation. Reader who is not fond of too much abstraction can skip these parts and absorb only the statement of the Laplace theorem and its corollaries.

The total fee for the entry into the watchtower and the botanical garden expressed in our variables gives $30x + 16y + 24z$ and $40x + 24y + 34z$ respectively. But we know the actual values (€1 116 and €1 542). Thus we have

$$\begin{aligned} 30x &+ 16y + 24z = 1\,116,\\ 40x &+ 24y + 34z = 1\,542. \end{aligned}$$

We write the system of three linear equations in the matrix notation as

$$\begin{pmatrix} 1 & 1 & 1\\ 30 & 16 & 24\\ 40 & 24 & 34 \end{pmatrix} \cdot \begin{pmatrix} x\\ y\\ z \end{pmatrix} = \begin{pmatrix} 45\\ 1\,116\\ 1\,542 \end{pmatrix}.$$

The solution is

$$\begin{pmatrix} x\\ y\\ z \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 16 & 5 & -4\\ 30 & 3 & -3\\ -40 & -8 & 7 \end{pmatrix} \cdot \begin{pmatrix} 45\\ 1\,116\\ 1\,542 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 132\\ 72\\ 66 \end{pmatrix} = \begin{pmatrix} 22\\ 12\\ 11 \end{pmatrix},$$

because

$$\begin{pmatrix} 1 & 1 & 1\\ 30 & 16 & 24\\ 40 & 24 & 34 \end{pmatrix}^{-1} = \frac{1}{6}\begin{pmatrix} 16 & 5 & -4\\ 30 & 3 & -3\\ -40 & -8 & 7 \end{pmatrix}.$$

Expressed in words, there were 22 adults, 12 children and 11 seniors.

□

2.27. Using the inverse matrix, compute the solution of the system

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 2,\\ x_1 + x_2 - x_3 - x_4 &= 3,\\ x_1 - x_2 + x_3 - x_4 &= 3,\\ x_1 - x_2 - x_3 + x_4 &= 5. \end{aligned}$$

○

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 then always has more than one solution. As the reader may know, a system of linear equations either has no solution, has one solution or has infinitely many solutions (for instance, it cannot have exactly two solutions). 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 *homogeneous system* of linear equations) or an affine space, see 4.1 (in the case when the right-hand side of at least one of the equations is non-zero, we speak of *non-homogeneous system* of linear equations). We demonstrate possible types of solutions of a system of linear equations by examples.

2.20. Minors of the matrix. When investigating matrices and their properties we often work only with parts of the matrices. Therefore we need some new notions.

SUBMATRICES AND MINORS

Let $A = (a_{ij})$ be a matrix of the type m/n and let $1 \le i_1 < \dots < i_k \le m$, $1 \le j_1 < \dots < j_l \le n$ be fixed natural numbers. Then the matrix

$$M = \begin{pmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \dots & a_{i_1 j_\ell}\\ \vdots & & & \vdots\\ a_{i_k j_1} & a_{i_k j_2} & \dots & a_{i_k j_\ell} \end{pmatrix}$$

of the type k/ℓ is called a *submatrix of the matrix A* determined by the rows $i_1, \dots, i_k$ and columns $j_1, \dots, j_\ell$. The remaining $(m - k)$ rows and $(n - l)$ 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 define $|M|$, which is called *subdeterminant* or *minor* of the order k of the matrix A. If $m = n$, then when $k = \ell$ we have also M^* square, then $|M^*|$ is called *minor complement* $|M|$, or *complementary minor* of the submatrix M in the matrix A. The scalar

$$(-1)^{i_1+\dots+i_k+j_1+\dots+j_l} \cdot |M^*|$$

is called *algebraic complement* of the minor $|M|$.

Submatrices formed by the first k rows and columns are called *leading principal submatrices*, and their determinants *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 matrices* and *principal minors*.

Specially, when $k = \ell = 1$, $m = n$ we call the corresponding complementary minor the *algebraic complement A_{ij}* of the element a_{ij} of the matrix A.

2.21. 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 we see that each of the individual $k!(n-k)!$ members in the product of $|M|$ with its algebraic complement is a member of $|A|$.

In general, consider a submatrix M, that is, a square matrix given by the rows $i_1 < i_2 < \dots < i_k$ and columns $j_1 < \dots < j_k$. Then using $(i_1 - 1) + \dots + (i_k - k)$ exchanges of neighbouring rows and $(j_1 - 1) + \dots + (j_k - k)$ exchanges of neighbouring columns in A we can transform this submatrix M into a principal submatrix and the complementary matrix gets transformed into its complementary matrix. The whole matrix A gets transformed into a matrix B for which it holds thanks to 2.17 and the definition of determinant that $|B| = (-1)^\alpha |A|$, where $\alpha = \sum_{h=1}^{k}(i_h - j_h) - 2(1 + \dots + k)$. 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 member of $|M|$ with any member of its algebraic complement is a member of $|A|$.*

This claim suggests the intuition than using some products of smaller determinants we could express the determinant of the matrix itself. We see that $|A|$ contains exactly $n!$ distinct members, exactly one for each permutation. These members are mutually distinct as polynomials in elements of (a general indeterminate)

2.28. 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?

Solution. We rewrite it, as usual, in the extended matrix, and transform:

$$\left(\begin{array}{ccc|c} 1 & -a & -2 & b \\ 1 & 1-a & 0 & b-3 \\ 1 & 1-a & a & 2b-1 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -a & -2 & b \\ 0 & 1 & 2 & -3 \\ 0 & 1 & a+2 & b-1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|c} 1 & -a & -2 & b \\ 0 & 1 & 2 & -3 \\ 0 & 0 & a & b+2 \end{array}\right).$$

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 $a \neq 0$. For $a = 0$, the third column is a zero column. If $a = 2$ and $b = -2$, we have a zero row, and choosing $x_3 \in \mathbb{R}$ as a parameter 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.

Let us 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

$$\left(\frac{-3a^2 - ab - 4a + 2b + 4}{a}, -\frac{2b+3a+4}{a}, \frac{b+2}{a}\right).$$

□

2.29. Determine the number of solutions for the systems

(a)

$$\begin{array}{rcrcrcl} 12x_1 & + & \sqrt{5}x_2 & + & 11x_3 & = & -9, \\ x_1 & & & - & 5x_3 & = & -9, \\ x_1 & & & + & 2x_3 & = & -7; \end{array}$$

(b)

$$\begin{array}{rcrcrcl} 4x_1 & + & 2x_2 & - & 12x_3 & = & 0, \\ 5x_1 & + & 2x_2 & - & x_3 & = & 0, \\ -2x_1 & - & x_2 & + & 6x_3 & = & 4; \end{array}$$

(c)

$$\begin{array}{rcrcrcl} 4x_1 & + & 2x_2 & - & 12x_3 & = & 0, \\ 5x_1 & + & 2x_2 & - & x_3 & = & 1, \\ -2x_1 & - & x_2 & + & 6x_3 & = & 0. \end{array}$$

matrix A. If we can show that there are exactly that many mutually distinct expressions from the previous claim, we obtain the determinant $|A|$ from their sum.

It remains to show that the members 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 $k!(n-k)!$ members in the products of $|M|$ with their algebraic complements is a member of $|A|$. But for distinct choices of M we can never obtain the same members and the individual members in $(-1)^{i_1+\dots+i_k+j_1+\dots+j_l} \cdot |M| \cdot |M^*|$ are also mutually distinct. Therefore we have exactly the required number $k!(n-k)!\binom{n}{k} = n!$ of members.

Thus we have proven:

Laplace theorem

Theorem. *Let $A = (a_{ij})$ be a square matrix of dimension n over arbitrary ring of scalars with k rows fixed. Then $|A|$ is a sum of all $\binom{n}{k}$ products $(-1)^{i_1+\dots+i_k+j_1+\dots+j_l} \cdot |M| \cdot |M^*|$ of minors of the order k chosen among the fixed rows with their algebraic complements.*

Laplace theorem transforms the computation of $|A|$ into the computation of determinants of lower dimension. This method of computation is called *Laplace expansion* by the chose rows (or columns). For instance, the expansion of the i-th row or j-th column is:

$$|A| = \sum_{j=1}^{n} a_{ij} A_{ij}$$

where A_{ij} denotes the algebraic complement of the element a_{ij} (that is, minor of order one).

In practical computation it is often efficient to combine the Laplace expansion with a direct method of linear combination addition.

2.22. Proof of Cauchy theorem. The theorem is based on a

clever but elementary application of the Laplace theorem. We just use the Laplace expansion twice on particular positions in the matrix.

Let us first consider the following matrix H of the 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).

$$H = \begin{pmatrix} A & 0 \\ -E & B \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} & 0 & \dots & 0 \\ -1 & & 0 & b_{11} & \dots & b_{1n} \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & b_{n1} & \dots & b_{nn} \end{pmatrix}$$

Laplace expansion of the first n rows gives us $|H| = |A| \cdot |B|$.

Now we add sequentially to the last n rows linear combinations of the first n columns in order to obtain a matrix with zeros in the

Solution. 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 is invertible. Thus the system for the case (a) has exactly one solution.

For the cases (b) and (c) it is enough to note that

$$(4, 2, -12) = -2(-2, -1, 6).$$

In the case (b) adding the first equation to the third multiplied by two gives $0 = 8$, no solution for the system; in the case (c) the third equation is a multiple of the first – the system has clearly infinitely many distinct solutions. □

2.30. Find (any) 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 exactly for every $t \in \mathbb{R}$ and vectors

$$(2, -1, 0, 0), \quad (0, 2, 0, -1), \quad (0, 0, 4, -3)$$

giving the left-hand sides of the equations are clearly linearly independent (the set of solutions contains a single parameter). □

2.31. 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

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & + & x_3 & - & 2x_4 & = & 4, \\ -3x_1 & - & 2x_2 & - & x_3 & - & x_4 & = & 5, \\ & + & 2x_2 & & & + & x_4 & = & 1, \\ x_1 & - & 4x_2 & + & x_3 & - & 2x_4 & = & 3 \end{array}$$

and all solutions of the system

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & + & x_3 & - & 2x_4 & = & 0, \\ -3x_1 & - & 2x_2 & - & x_3 & - & x_4 & = & 0, \\ & + & 2x_2 & & & + & x_4 & = & 0, \\ x_1 & - & 4x_2 & + & x_3 & - & 2x_4 & = & 0 \end{array}$$

and of the system

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & & & = & 1, \\ x_1 & - & 2x_2 & + & 2x_3 & = & -4, \\ x_1 & - & x_2 & & & = & 1, \\ -2x_1 & - & x_2 & + & x_3 & = & -2. \end{array}$$

bottom right corner. We obtain

$$K = \begin{pmatrix} a_{11} & \dots & a_{1n} & c_{11} & \dots & c_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} & c_{n1} & \dots & c_{nn} \\ -1 & & 0 & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & 0 & \dots & 0 \end{pmatrix}.$$

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 members of the product $A{\cdot}B$ and $|K| = |H|$. The expansion of the last n columns gives us

$$|K| = (-1)^{n+1+\dots+2n}|A \cdot B| = (-1)^{2n\cdot(n+1)} \cdot |A \cdot B| = |A \cdot B|.$$

This proves the Cauchy theorem.

2.23. 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 $|E| = 1$, for every invertible matrix we have that $|A|$ is an invertible scalar and thanks to the Cauchy theorem we have $|A^{-1}| = |A|^{-1}$.

But we can say more, combining the Laplace and Cauchy theorem.

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 *algebraically adjoint matrix* of the matrix A.

Theorem. *For every square matrix A over a ring of scalars $\mathbb{K}$ we have that*

$$AA^* = A^*A = |A| \cdot E. \tag{2.2}$$

Notably,

(1) A^{-1} exists as a matrix over a ring of scalars $\mathbb{K}$ if and only if $|A|^{-1}$ exists in $\mathbb{K}$.
(2) If A^{-1} exists, then $A^{-1} = |A|^{-1} \cdot A^*$.

Proof. As we have already mentioned, Cauchy theorem shows that the existence of A^{-1} implies the invertibility of $|A| \in \mathbb{K}$.

For 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 $|A|$ of i-th row. If $i \neq j$ it is expansion of determinant of the 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 the equality (2.2).

Let us further assume that $|A|$ is an invertible scalar. If we repeat the previous computation for $A^*{\cdot}A$, we obtain $|A|^{-1}A^*{\cdot}A = E$. Therefore our computation really gives the inverse matrix of A, as claimed in the theorem. □

Solution. Because $\det A = -10$, that is, non-zero, the columns of A are linearly independent, and thus the rank equals to the dimension.

The first of the three given system is given by the extended matrix

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & -2 & 4 \\ -3 & -2 & -1 & -1 & 5 \\ 0 & 2 & 0 & 1 & 1 \\ 1 & -4 & 1 & -2 & 3 \end{array}\right).$$

But the left-hand side is exactly A^T with determinant $|A^T| = |A| \neq 0$. 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

$$(x_1,\, x_2,\, x_3,\, x_4) = (0, 0, 0, 0) .$$

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

$$\left(\begin{array}{ccc|c} 0 & 0 & 0 & a \end{array}\right), \quad \text{kde} \quad a \neq 0.$$

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. □

2.32. Let

$$A = \begin{pmatrix} 4 & 5 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

be given. 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.

For the readers it is definitely no problem to find correct values in the cases a) and c) (it is enough to choose $b_1 = b_2 + b_3$ in the case

As a direct corollary of this theorem we can once again prove the Cramer rule for solving the systems of linear equations, see 2.18. 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 last expression as the Laplace expansion 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

2.24. Abstract vector spaces. Let us go back for a while to the systems of m linear equations of n variables from 2.3 and let us further assume that the system is homogeneous $A \cdot x = 0$, that is,

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Thanks to 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 is thus 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 on vector addition and scalar multiplication. That are the basic properties of vectors of dimension n in $\mathbb{K}^n$, see 2.1. Now we have the vectors in the solution space with n coordinates and 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 have with the solution of 1000 coordinates 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 1.25 at the page 25.

Already in the paragraph 1.9 we have encountered a more 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. Therefore we need a more general definition of vector space and its dimension:

Vector space definition

Vector space V over field of scalars $\mathbb{K}$ is a set where we define the operations

- addition, which satisfies the axioms (KG1)–(KG4) from the paragraph 1.1 on the page 4,
- scalar multiplication, for which the axioms axioms (V1)–(V4) from the paragraph 2.1 on the page 72 hold.

Let us remind our simple notational convention: scalars are usually denoted by letters from the beginning of the alphabet, that is, $a, b, c, \dots$, while for vectors we shall use letters from the end, that is, u, v, w, x, y, z. Usually, x, y, z will denote n-tuples of

a) and $b_1 \neq b_2 + b_3$ in the case c)). Let us further note that $|A| = 0$, thus the system either has infinitely many or no solution. In general, the set of solutions of a homogeneous system of linear equations is a vector space, thus the variant d) is a priori excluded. The variant b) is possible only for a system with a regular matrix (the only solution is then a zero vector).

□

2.33. Solve the system of homogeneous linear equations given by the matrix

$$\begin{pmatrix} 0 & \sqrt{2} & \sqrt{3} & \sqrt{6} & 0 \\ 2 & 2 & \sqrt{3} & -2 & -\sqrt{5} \\ 0 & 2 & \sqrt{5} & 2\sqrt{3} & -\sqrt{3} \\ 3 & 3 & \sqrt{3} & -3 & 0 \end{pmatrix}.$$

○

2.34. Determine all solutions of the system

$$\begin{array}{rcrcrcrcr} & & x_2 & & & + & x_4 & = & 1, \\ 3x_1 & - & 2x_2 & - & 3x_3 & + & 4x_4 & = & -2, \\ x_1 & + & x_2 & - & x_3 & + & x_4 & = & 2, \\ x_1 & & & - & x_3 & & & = & 1. \end{array}$$

○

2.35. Solve

$$\begin{array}{rcrcrcrcr} 3x & - & 5y & + & 2u & + & 4z & = & 2, \\ 5x & + & 7y & - & 4u & - & 6z & = & 3, \\ 7x & - & 4y & + & & + & 3z & = & \end{array}$$

○

2.36. Decide whether the system of linear equations

$$\begin{array}{rcrcrcr} 3x_1 & + & 3x_2 & + & x_3 & = & 1, \\ 2x_1 & + & 3x_2 & - & x_3 & = & 8, \\ 2x_1 & - & 3x_2 & + & x_3 & = & 4, \\ 3x_1 & - & 2x_2 & + & x_3 & = & 6 \end{array}$$

of three variables x_1, x_2, x_3 is solvable. ○

2.37. Determine the number of solutions of 2 systems of 5 linear equations

$$A^T \cdot x = (1, 2, 3, 4, 5)^T, \quad A^T \cdot x = (1, 1, 1, 1, 1)^T$$

where

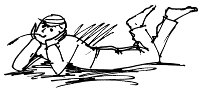
$$x = (x_1, x_2, x_3)^T \quad \text{a} \quad A = \begin{pmatrix} 3 & 1 & 7 & 5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 2 & 1 & 4 & 3 & 0 \end{pmatrix}.$$

○

2.38. Determine the solution of the system of linear equations

$$\begin{array}{rcrcrcr} ax_1 & + & 4x_2 & +2 & x_3 & = & 0, \\ 2x_1 & + & 3x_2 & - & x_3 & = & 0, \end{array}$$

scalars. For completeness, the letters from the middle of the alphabet, for instance i, j, k, ℓ, will mostly denote indices in expressions.



In order to gain some practice in formal approach, we check simple properties of vectors, which are trivial for n-tuples for scalars, but not so evident for general vectors.

2.25. Proposition. *Let V be a vector space over a field of scalars $\mathbb{K}$, further take $a, b, a_i \in \mathbb{K}$, and vectors $u, v, u_j \in V$. 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

$$(a + 0) \cdot u \overset{(V2)}{=} a \cdot u + 0 \cdot u = a \cdot u$$

which according to the axiom (KG4) ensures $0 \cdot u = 0$. Now

$$u + (-1) \cdot u \overset{(V2)}{=} (1 + (-1)) \cdot u = 0 \cdot u = 0$$

and thus $-u = (-1) \cdot u$. Further,

$$a \cdot (u + (-1) \cdot v) \overset{(V2,V3)}{=} a \cdot u + (-a) \cdot v = a \cdot u - a \cdot v,$$

Which proves (3). It holds that

$$(a - b) \cdot u \overset{(V2,V3)}{=} a \cdot u + (-b) \cdot u = a \cdot u - b \cdot u$$

which proves (4). The 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 first need the axiom of the field for scalars and the axiom (V4) for vector spaces: if $p \cdot u = 0$ and $p \neq 0$, then $u = 1 \cdot u = (p^{-1} \cdot p) \cdot u = p^{-1} \cdot 0 = 0$. □

2.26. Linear (in)dependence. In the paragraph 2.11 we have worked with the so-called linear combinations of rows of a matrix. With general vectors we will work analogously:

Linear combinations and independence

Expression of the form $a_1 \cdot v_1 + \cdots + a_k \cdot v_k$ is called *linear combination* of vectors $v_1, \dots, v_k \in V$.

Finite sequence of vectors $v_1, \dots, v_k$ is called *linearly independent*, if the only zero linear combination is the one with all coefficients zero, that is, for scalars $a_1, \dots, a_k \in \mathbb{K}$ holds that

$$a_1 \cdot v_1 + \cdots + a_k \cdot v_k = 0 \implies a_1 = a_2 = \cdots = a_k = 0.$$

It is clear that in independent sequence of vectors all vectors are mutually distinct and nonzero.

The set of vectors $M \subset V$ in vector space V over $\mathbb{K}$ is called *linearly independent*, if every finite k-tuple of vectors $v_1, \dots, v_k \in M$ is linearly independent.

The set of vectors M is *linearly dependent*, if it is not linearly independent.

depending on the parameter $a \in \mathbb{R}$. ○

2.39. Depending on the parameter $a \in \mathbb{R}$ determine the number of solutions of the system

$$\begin{pmatrix} 4 & 1 & 4 & a \\ 2 & 3 & 6 & 8 \\ 3 & 2 & 5 & 4 \\ 6 & -1 & 2 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 3 \\ -3 \end{pmatrix}.$$

○

2.40. Decide whether 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) $\{(x, y, 2y);\ x, y \in \mathbb{R}\}$.

○

2.41. Solve the system of linear equations, depending on the real parameters a, b.

$$\begin{aligned} x + 2y + bz &= a \\ x - y + 2z &= 1 \\ 3x - y &= 1. \end{aligned}$$

○

2.42. 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

$$A_{11} = \begin{vmatrix} 3 & 0 & 4 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} = -24, \qquad A_{12} = -\begin{vmatrix} 0 & 0 & 4 \\ 5 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = 0, \dots$$

$$A_{43} = -\begin{vmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 0 \end{vmatrix} = 0, \qquad A_{44} = \begin{vmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 5 & 0 & 6 \end{vmatrix} = -12.$$

Directly from the definition we have that a nonempty subset M of vectors from 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. Really, at least one of the coefficients in the corresponding zero 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 using 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.

2.27. Generators and subspaces. Subset $M \subset V$ is called *vector subspace* if it along with restricted operations of addition and scalar multiplication forms a vector space. That is, we require

$$\forall a, b \in \mathbb{K},\ \forall v, w \in M,\ a \cdot v + b \cdot w \in M.$$

Let us 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" one-dimensional subspace, while over $\mathbb{Q}$ the subspace is "bigger".

Polynomials of degree at most m form a vector space $\mathbb{R}_m[x]$. We can see the polynomials as mappings $f : \mathbb{R} \to \mathbb{R}$ and define the addition and scalar multiplication like this: $(f + g)(x) = f(x) + g(x)$, $(a \cdot f)(x) = a \cdot f(x)$. Polynomials of any degree also form a vector space $\mathbb{R}_\infty[x]$ and $\mathbb{R}_m[x] \subset \mathbb{R}_n[x]$ is a vector subspace for any $m \le n \le \infty$. Subspaces are also for instance all even polynomials or odd polynomials, that is, polynomials satisfying $f(-x) = \pm f(x)$.

In a complete analogy as with polynomials we can define a structure of vector space 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 clearly see it 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 for all $i \in I$, $a \cdot u + b \cdot v \in W_i$, but that means that $a \cdot u + b \cdot v \in \cap_{i \in I} W_i$.

Notably, the intersections $\langle M \rangle$ of all subspaces $W \subset V$ that contain some given set of vectors $M \subset V$ is a subspace.

We say that a set M *generates* the subspace $\langle M \rangle$, or that the elements of M are *generators* of the subspace $\langle M \rangle$.

Let us again formulate a few simple claims about subspace generation:

Proposition. *For every nonempty set $M \subset V$ we have that*

By plugging in 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 $A^{-1} = |A|^{-1} \cdot A^*$. 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 plugging in we obtain

$$A^{-1} = \begin{pmatrix} -3/2 & 0 & 1/2 & 0 \\ 0 & -2 & 0 & 1 \\ 5/4 & 0 & -1/4 & 0 \\ 0 & 7/4 & 0 & -3/4 \end{pmatrix}.$$

□

2.43. Find the algebraically adjoint matrix F^* for

$$F = \begin{pmatrix} \alpha & \beta & 0 \\ \gamma & \delta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \alpha, \beta, \gamma, \delta \in \mathbb{R}.$$

○

2.44. Calculate the algebraically adjoint matrix for the matrices

$$\text{(a)} \begin{pmatrix} 3 & -2 & 0 & -1 \\ 0 & 2 & 2 & 1 \\ 1 & -2 & -3 & -2 \\ 0 & 1 & 2 & 1 \end{pmatrix}, \qquad \text{(b)} \begin{pmatrix} 1+i & 2i \\ 3-2i & 6 \end{pmatrix},$$

where i denotes the imaginary unit.

○

F. Vector spaces

The properties of vector space, which we have already observed for the plane or three dimensional space are possessed by other sets as well. We illustrate this by examples.

2.45. Vector space – yes or no? Decide for the following sets whether they form a vector space over the field of real numbers:

(1) $\langle M\rangle = \{a_1 \cdot u_1 + \cdots + a_k \cdot u_k;\ k \in \mathbb{N}, a_i \in \mathbb{K}, u_j \in M, j = 1, \dots, k\}$;

(2) $M = \langle M\rangle$ if and only if M is a vector subspace;

(3) if $N \subset M$ then $\langle N\rangle \subset \langle M\rangle$ is a vector subspace

Subspace $\langle\emptyset\rangle$ generated by the empty subspace is the trivial subspace $\{0\} \subset V$.

Proof. (1) The set of all linear combinations

$$a_1u_1 + \cdots + a_ku_k$$

on the right-hand side (1) is clearly a vector subspace and of course it contains M. On the other hand, each of the linear combinations must be in $\langle M\rangle$ and thus the first claim is proven.

The claim (2) follows immediately from (1) and from the definition of vector space and analogously (1) implies the third claim.

Finally, the smallest subspace is $\{0\}$, because empty set is contained in every subspace and each of them contains the vector 0. □

2.28. Sums of subspace. Since we now have some intuition about generators and their respective subspaces, we should understand the possibilities how some subspaces can generate 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, $\langle \cup_{i\in I} V_i\rangle$, is called *sum of subspaces* V_i. We denote it as $\sum_{i\in I} V_i$. Notably, for a finite number of subspaces $V_1, \dots, V_k \subset V$ we write

$$V_1 + \cdots + V_k = \langle V_1 \cup V_2 \cup \cdots \cup V_k\rangle.$$

We see that every element in the considered subspace can be expressed as a linear combination of vectors from the subspaces V_i. Because vector addition is commutative, we can associate members that belong to the same subspace and for a finite sum of k subspaces we obtain

$$V_1 + V_2 + \cdots + V_k = \{v_1 + \cdots + v_k;\ v_i \in V_i, i = 1, \dots, k\}.$$

Sum $W = V_1 + \cdots + V_k \subset V$ is called *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 case can every vector $w \in W$ be written in a unique way as a sum

$$w = v_1 + \cdots + v_k,$$

where $v_i \in V_i$. Really, if for that vector we could simultaneously write $w = v_1' + \cdots + v_k'$, 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 other subspaces. That is a contradiction with the assumption that V_i has zero intersection with 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

$$W = V_1 \oplus \cdots \oplus V_k = \oplus_{i=1}^k V_i.$$

2.29. Basis. Now we have everything prepared for understanding minimal sets of generators as we understood them in the plane $\mathbb{R}^2$.

i) The set of solutions of the system

$$\begin{aligned} x_1+x_2+\cdots+x_{98}+x_{99}+x_{100} &= 100x_1,\\ x_1+x_2+\cdots+x_{98}+x_{99} &= 99x_1,\\ x_1+x_2+\cdots+x_{98} &= 98x_1,\\ &\vdots\\ x_1+x_2 &= 2x_1. \end{aligned}$$

ii) The set of solutions of the equation

$$x_1+x_2+\cdots+x_{100}=0$$

iii) The set of solution 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 homogeneous difference equation.

vi) The set of solutions of non-homogeneous difference equation.

vii) $\{f:\mathbb{R}\to\mathbb{R}|f(1)=f(2)=c, c\in\mathbb{R}\}$

Solution.

i) Yes. They all are real multiples of the vector $\underbrace{(1,1,1\dots,1)}_{\text{100 ones}}$, that is, vector space of dimension 1 (see also (2.29)).

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 homogeneous linear equations forms a vector space.

iii) No. For instance, taking twice the solution $x_1=1$, $x_i=0$, $i=2,\dots 100$ we do not obtain a solution. But the set of solutions forms a so-called affine space (see (4.1)).

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). Let us have two sequences $(x_j)_{j=0}^{\infty}$ and $(y_j)_{j=0}^{\infty}$

Basis of vector space

Subset $M\subset V$ is called *basis of vector space* V if $\langle M\rangle=V$ and M is linearly independent.

Vector space with 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 $k=\infty$.

In order to be satisfied with such definition of dimension, we must know that different bases of the same space will always have the same number of elements. This we will show in a while. But we note immediately, that the trivial subspace is generated by empty set, which is an "empty" basis. It thus has zero dimension.

Basis of a k-dimensional space will usually be denoted as a k-tuple $\underline{v}=(v_1\dots,v_k)$ of basis vectors. It is mostly about having a convention: with finitely dimensional vector spaces we shall always consider the base along with a given order of the elements even if we have not defined it that way, strictly said.

Clearly, if $(v_1,\dots,v_n)$ is a basis of V, the whole space V is a direct sum of the one-dimensional subspaces

$$V=\langle v_1\rangle\oplus\cdots\oplus\langle v_n\rangle.$$

An immediate corollary of the derived uniqueness of decomposition of any vector w in V into the components in the direct sum gives unique decomposition

$$w=x_1v_1+\cdots+x_nv_n$$

and allows us after choosing a basis to see vectors again as n-tuples of scalars. To this idea we will return in the paragraph 2.33, when we finish the discussion of existence of bases and sums of subspaces in general case.

2.30. Theorem. >*From any finite set of generators of a vector space V we can choose a basis. Every base of a finitely dimensional space V has the same number of elements.*

Proof. First claim can be easily proved using induction on the number of generators k.

Only the zero subspace does not need any generator and thus we are able to choose an empty basis. On the other hand, we are not allowed to choose the zero vector (generators would 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 $V=\langle\{v\}\rangle$ and $v\neq 0$, because $\{v\}$ is linearly independent set of vectors. Then $\{v\}$ is also a basis of the vector space V.

Assume that the claim holds for $k=n$ and consider $V=\langle v_1,\dots,v_{n+1}\rangle$. If $v_1,\dots,v_{n+1}$ are linearly independent, then they form a basis. In the other case there exists i such that

$$v_i=a_1v_1+\cdots+a_{i-1}v_{i-1}+a_{i+1}v_{i+1}+\cdots+a_{n+1}v_{n+1}.$$

Then $V=\langle v_1,\dots,v_{i-1},v_{i+1},\dots,v_{n+1}\rangle$ and we can choose a basis, using inductive assumption.

In remains to ensure that bases always have the same number of elements. Consider basis $(v_1,\dots,v_n)$ of the space V and for arbitrary nonzero vector consider

$$u=a_1v_1+\cdots+a_nv_n\in V$$

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

$$\begin{aligned} a_n x_{n+k} + a_{n-1} x_{n+k-1} + \cdots + a_0 x_k &= c \\ a_n y_{n+k} + a_{n-1} y_{n+k-1} + \cdots + a_0 y_k &= c, \quad c \in \mathbb{R} - \{0\} \end{aligned}$$

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.

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 $2c = c$, therefore $c = 0$.

□

2.46. 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 set U_1 is not a vector (sub)space. We can see that, for instance,

$$(1, 1, 1) + (-1, 1, 1) = (0, 2, 2) \notin U_1.$$

The set U_2 is a subspace (there is a clear identification with $\mathbb{R}^2$), because

$$\left(a_1 x^2 + c_1\right) + \left(a_2 x^2 + c_2\right) = (a_1 + a_2)\, x^2 + (c_1 + c_2),$$

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 $\langle u, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n \rangle = V$.

We ensure that this is again a basis: if adding u to linearly independent vectors $v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n$ would lead to a set of linearly dependent vectors, then u is their linear combination. That would mean

$$V = \langle v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n \rangle,$$

which is not possible.

Thus we have proved that for any nonzero vector $u \in V$ there exists i, $1 \leq i \leq n$, such that $(u, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n)$ is again a basis of V.

Further, we shall instead of one vector u consider a linearly independent set $u_1, \dots, u_k$ and we will sequentially add $u_1, u_2, \dots,$ always exchanging for some v_i using our previous approach. We have to ensure that there always is such v_i (that is, that the vector u will not exchange for each other). Assume thus that we have already placed $u_1, \dots, u_\ell$. Then the vector $u_{\ell+1}$ can clearly be expressed as a linear combination of such vector and the remaining v_j. If only the coefficients at $u_1, \dots, u_\ell$ were nonzero, that would mean that the vectors $u_1, \dots, u_{\ell+1}$ are linearly dependent, which is a contradiction.

For every $k \leq n$ we can after k steps obtain a basis in which from the original basis k vectors were exchanged for new ones. If $k > n$, then in the n-th step we obtain a basis consisting only of new vectors u_i, which means that the original set could not be linearly independent. Notably it is not possible that two bases have a different number of elements. □

In reality, we have proved a stronger claim, the so-called *Steinitz exchange theorem*, which says that for every finite basis $\underline{v}$ and every system of linearly independent vectors in V we can find a subset of the basis vectors v_i which can be exchanged with the new vectors to obtain a basis.

2.31. Corollaries of the Steinitz exchange theorem. Thanks to the possibility of freely choosing and exchanging basis vectors we can immediately derive nice (and intuitively expectable) properties of bases of vector space:

Proposition. *(1) Every two bases of a finitely 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) Basis of a finitely dimensional vector space is maximal linearly independent set.

(4) Bases of a vector space are exactly minimal sets of generators.

A little bit 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 we have that*

(1) $\dim W \leq \dim V$,

(2) $V = W$ *if and only if* $\dim V = \dim W$,

(3) $\dim W_1 + \dim W_2 = \dim(W_1 + W_2) + \dim(W_1 \cap W_2)$.

$$k \cdot \left(ax^2 + c\right) = (ka)\,x^2 + kc$$

for all numbers $a_1, c_1, a_2, c_2, a, c, k \in \mathbb{R}$. □

2.47. 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? ○

G. Linear dependence and independence, bases

2.48. By calculating the determinant of a suitable matrix decide whether the vectors $(1, 2, 3, 1)$, $(1, 0, -1, 1)$, $(2, 1, -1, 3)$ and $(0, 0, 3, 2)$ are linearly dependent or not.

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. □

2.49. Given arbitrary linearly independent vectors u, v, w, z in a vector space V, decide whether in V the vectors

$$u - 2v, \quad 3u + w - z, \quad u - 4v + w + 2z, \quad 4v + 8w + 4z$$

are linearly independent or not.

Solution. The 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. □

2.50. 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 rank lower than the number of the vectors, which in this case means that rank must be two or lower. In

Proof. It remains to prove only the last claim. That is clear when the dimension of one of the spaces is zero. Assume then that $\dim W_1 = r \geq 1$, $\dim W_2 = s \geq 1$ and let $(w_1 \dots, w_t)$ be a basis of $W_1 \cap W_2$ (or empty set, if the intersection is trivial).

According to the Steinitz exchange theorem this basis of the intersection can be extended to a basis $(w_1, \dots, w_t, u_{t+1} \dots, u_r)$ for W_1 and to a basis $(w_1 \dots, w_t, v_{t+1}, \dots, v_s)$ for W_2. Vectors

$$w_1, \dots, w_t, u_{t+1}, \dots, u_r, v_{t+1} \dots, v_s$$

clearly generate $W_1 + W_2$. We show that they are linearly independent. Let

$$\begin{aligned} &a_1 w_1 + \dots + a_t w_t + b_{t+1} u_{t+1} + \dots \\ &\qquad \dots + b_r u_r + c_{t+1} v_{t+1} + \dots + c_s v_s = 0. \end{aligned}$$

Then necessarily

$$\begin{aligned} &-(c_{t+1} \cdot v_{t+1} + \dots + c_s \cdot v_s) = \\ &\quad = a_1 \cdot w_1 + \dots + a_t \cdot w_t + b_{t+1} \cdot u_{t+1} + \dots + b_r \cdot u_r \end{aligned}$$

must belong to $W_2 \cap W_1$. That implies that

$$b_{t+1} = \dots = b_r = 0,$$

since in that way we have defined our bases. Then also

$$a_1 \cdot w_1 + \dots + a_t \cdot w_t + c_{t+1} \cdot v_{t+1} + \dots + c_s \cdot v_s = 0$$

and because the corresponding vectors form a basis W_2, all the coefficients are zero.

The claim (3) now follows by directly calculating of generators. □

2.32. Examples. (1) $\mathbb{K}^n$ has (as a vector space over $\mathbb{K}$) dimension n. Basis is for example an n-tuple of vectors

$$((1, 0, \dots, 0), (0, 1, \dots, 0) \dots, (0, \dots, 0, 1)).$$

This basis is called the *standard basis of* $\mathbb{K}^n$. Note that in the case of finite field of scalars, say $\mathbb{Z}_k$, 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, basis is for instance the numbers 1 and i.

(3) $\mathbb{K}_m[x]$, that is, the space of all polynomials of degree at most m, has dimension $m + 1$, basis is for instance the sequence $1, x, x^2, \dots, x^m$.

Vector space of all polynomials $\mathbb{K}[x]$ has dimension ∞, but we can still find a basis (although infinite in size): $1, x, x^2, \dots$.

(4) Vector space $\mathbb{R}$ over $\mathbb{Q}$ has dimension ∞ and does not have a countable basis.

(5) Vector space of all mappings $f : \mathbb{R} \to \mathbb{R}$ has also dimension ∞ and does not have any finite basis.

2.33. Vector coordinates. If we fix a basis $(v_1, \dots, v_n)$ of a finitely dimensional space V, then every vector $w \in V$ can be expressed as a linear combination $v = a_1 v_1 + \dots + a_n v_n$. Assume that we can do it in two ways:

$$w = a_1 v_1 + \dots + a_n v_n = b_1 v_1 + \dots + b_n v_n.$$

But then

$$0 = (a_1 - b_1) \cdot v_1 + \dots + (a_n - b_n) \cdot v_n$$

the case of square matrix, rank lower 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 $a_1 = -1$, $a_{2,3} = \frac{1 \pm \sqrt{21}}{2}$. □

2.51. 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:

$$\begin{aligned} a - b \quad &= 1 \\ 2a + b + c &= 1 \\ a + c &= 1, \end{aligned}$$

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 $(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$. □

2.52. 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

$$(5, 1, 11) = p\,(3, 2, 2) + q\,(2, 3, 1) + r\,(1, 1, 3)\,.$$

○

2.53. 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

and thus $a_i = b_i$ for all $i = 1, \dots, n$. We have reached the following conclusion:

In a finitely dimensional space every vector can be given in a unique way as a linear combination of basis vectors. Coefficients of this unique linear combination expressing the given vector $w \in V$ in the chosen basis $\underline{v} = (v_1, \dots, v_n)$ are called *coordinates of the vector* w in this basis.

Whenever we speak about coordinates $(a_1, \dots, a_n)$ of vector w, which we express as a sequence, we must have a fixed ordering of basis vectors $\underline{v} = (v_1, \dots, v_n)$. Although we have defined the basis as a minimal set of generators, in reality we work with them as with sequences (that is, with ordered sets).

Assigning coordinates to vectors

Mapping, which to the vector $u = a_1 v_1 + \cdots + a_n v_n$ assigns its coordinates in the basis $\underline{v}$ shall be denoted with 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 over left and right side of these equations are not identical, quite the opposite, they are operations over different vector spaces! At this opportunity, we can think about the general case of the basis M of (possibly infinite) vector space V. The basis then does not have to be countable, but still we can define the mapping $\underline{M} : V \to \mathbb{K}^M$ (that is, the coordinates of the vectors are the mapping from M to $\mathbb{K}$).

The given properties of assignments of coordinates were already seen at the mappings in geometry we have called linear (they preserved our linear structure in the plane). Before we deal more thoroughly with the dependency of the coordinates on the choice of the basis, we look in more generality at the notion of the linearity of the mapping.

2.34. Linear mapping. For any vector space (of finite or infinite dimension) we define "linearity" of a mapping between spaces similarly to the planar case ($\mathbb{R}^2$):

Linear mapping, definition

Let V and W be vector spaces over the same field of scalars $\mathbb{K}$. The mapping $f : V \to W$ is called *linear mapping (homomorphism)* if the following holds:

(1) $f(u + v) = f(u) + f(v)$, $\forall u, v \in V$
(2) $f(a \cdot u) = a \cdot f(u)$, $\forall a \in \mathbb{K}$, $\forall u \in V$.

Clearly, such mapping have already been seen in the case of matrix multiplication:

$$f : \mathbb{K}^n \to \mathbb{K}^m, x \mapsto A \cdot x$$

with matrix of type m/n over $\mathbb{K}$.

Image $\operatorname{Im} f := f(V) \subset W$ is always vector subspace, since linear combination of images $f(u_i)$ is an image of a 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 *kernel of linear mapping* f.

Linear mapping which is a bijection is called *isomorphism*.

$2+i$ 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

$$a\cdot(-\frac{1}{2}+i\frac{\sqrt{3}}{2})+b\cdot(-\frac{1}{2}-i\frac{\sqrt{3}}{2})=2+i.$$

By considering individually the real and the imaginary part of the equation we obtain a system of two linear equations in two variables:

$$\begin{aligned} -\frac{1}{2}a-\frac{1}{2}b &= 2 \\ \frac{\sqrt{3}}{2}a-\frac{\sqrt{3}}{2}b &= 1. \end{aligned}$$

Its solution gives us $a=-2+\frac{\sqrt{3}}{3}$, $b=-2-\frac{\sqrt{3}}{3}$, therefore the coordinates are $(-2+\frac{1}{\sqrt{3}}, -2-\frac{1}{\sqrt{3}})$. □

2.54. Remark. As a perceptive reader has definitely 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.

Let us also add a remark about the so-called 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.

2.55. 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.

2.56. 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?

2.57. Let a vector space V be given along with some basis formed by the vectors u, v, w, z. Determine whether the vectors

$$u-3v+z, \quad v-5w-z, \quad 3w-7z, \quad u-w+z$$

are linearly (in)dependent.

2.58. Complete the vectors $1-x^2+x^3$, $1+x^2+x^3$, $1-x-x^3$ into a basis of the space of polynomials of degree at most 3.

2.59. 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 matrix?

Analogously to the abstract definition of vector spaces, it is again necessary to prove seemingly trivial claims that follow from the axioms:

Proposition. *Let $f: V\to W$ be a linear mapping between two vector spaces over the same field of scalars $\mathbb{K}$. For all vectors $u, u_1, \dots, u_k \in V$ and scalars $a_1, \dots, a_k \in \mathbb{K}$ it holds that*

(1) $f(0)=0$,
(2) $f(-u)=-f(u)$,
(3) $f(a_1\cdot u_1+\dots+a_k\cdot u_k)=a_1\cdot f(u_1)+\dots+a_k\cdot f(u_k)$,
(4) for every vector subspace $V_1\subset V$ is its image $f(V_1)$ a vector subspace in W,
(5) for every vector subspace $W_1\subset W$ is the set $f^{-1}(W_1)=\{v\in V;\ f(v)\in W_1\}$ 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!):

$$\begin{aligned} f(0)=f(u-u)=f((1-1)\cdot u)=0\cdot f(u)=0, \\ f(-u)=f((-1)\cdot u)=(-1)\cdot f(u)=-f(u). \end{aligned}$$

The property (3) is again easy from the definition for two summands using induction on the number of summands. >From the property (3) we have that $\langle f(V_1)\rangle = f(V_1)$, thus it is vector subspace.

On the other hand, if $f(u)\in W_1$ and $f(v)\in W_1$ then for any scalars it will be that $f(a\cdot u+b\cdot v)=a\cdot f(u)+b\cdot f(v)\in W_1$. □

2.35. Simple corollaries.

(1) 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) Linear mapping $f: V\to W$ is an isomorphism if and only if $\operatorname{Im} f = W$ and $\operatorname{Ker} f=\{0\}\subset V$. Inverse mapping of an isomorphism is again an isomorphism.
(3) For any two subspaces $V_1, V_2\subset V$ and linear mapping $f: V\to W$ it holds that

$$\begin{aligned} f(V_1+V_2)&=f(V_1)+f(V_2),\\ f(V_1\cap V_2)&\subset f(V_1)\cap f(V_2). \end{aligned}$$

(4) The mapping "coordinate assignment" $\underline{u}: V\to\mathbb{K}^n$ given by arbitrarily chosen basis $\underline{u}=(u_1,\dots,u_n)$ of a vector space V is an isomorphism.
(5) Two finitely dimensional vector spaces are isomorphic if and only if they have the same dimension.
(6) Composition of two isomorphisms is an isomorphism.

Proof. Proving the first claim is a very easy exercise.

For the proof of the second one we must realise that if f is a linear bijection, then a vector w is an image of a linear combination $au+bv$, that is $w=f^{-1}(au+bv)$, if and only if

$$f(w)=au+bv=f(a\cdot f^{-1}(u)+b\cdot f^{-1}(v)).$$

Thus it also holds that $w=af^{-1}(u)+bf^{-1}(v)$ and therefore the inversion of a linear bijection is again a linear bijection.

Further, f is surjective if and only if $\operatorname{Im} f=W$ and 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.

The remaining claims are easy to prove by induction. Try to make a counterexample – in the inclusion that is to be proved there

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

$$\begin{pmatrix} 1 & 1 & -5 & 1 \\ 0 & 4 & 0 & -2 \\ 1 & 0 & 3 & 0 \\ -2 & -1 & 0 & 3 \end{pmatrix}$$

is regular (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 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). □

2.60. Let there be in $\mathbb{R}^3$ two vector spaces U and V generated by the vectors

$$(1, 1, -3)\,,\ (1, 2, 2) \quad \text{a} \quad (1, 1, -1)\,,\ (1, 2, 1)\,,\ (1, 3, 3)\,,$$

respectively. Determine the intersection of these two subspaces.

Solution. The subspace V has dimension only 2 (it is not the whole space $\mathbb{R}^3$), because

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ -1 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \end{vmatrix} = 0$$

and any two of the considered three vectors is clearly linearly independent. Similarly we can see that U has dimension 2. Also we have

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ -3 & 2 & -1 \end{vmatrix} = 2 \neq 0,$$

and therefore the vector $(1, 1, -1)$ does not lie in the subspace U. The intersection of two planes (two-dimensional spaces) passing through the origin in a three-dimensional space must be at least a line. In our case it is exactly a line (subspaces are not identical). Thus we have determined the dimension of the intersection – it is one-dimensional. If we note that

$$1 \cdot (1, 1, -3) + 2 \cdot (1, 2, 2) = (3, 5, 1) = 1 \cdot (1, 1, -1) + 2 \cdot (1, 2, 1)\,,$$

we obtain expression of the intersection in the form of a set of all scalar multiples of the vector $(3, 5, 1)$ (thus it is a line passing through the origin with this vector as a direction). □

does not always have to hold an equality (that is, find an example where the inclusion is strict). □

2.36. 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 $\underline{u} = (u_1, \dots, u_n)$ on V, $\underline{v} = (v_1, \dots, v_n)$ on W we have at our disposal the corresponding coordinate assignments and the whole situation is captured in the following diagram:

$$\begin{tikzcd} V \arrow[r, "f"] \arrow[d, "\underline{u}"', "\simeq"] & W \arrow[d, "\underline{v}", "\simeq"'] \\ \mathbb{K}^n \arrow[r, dotted, "f_{\underline{u},\underline{v}}"] & \mathbb{K}^m \end{tikzcd}$$

The bottom arrow $f_{\underline{u},\underline{v}}$ is defined by the remaining three, that is, as a mapping it is a composition

$$f_{\underline{u},\underline{v}} = \underline{v} \circ f \circ \underline{u}^{-1}.$$

Matrix of a linear mapping

Every linear mapping is uniquely determined by its values on an arbitrary set of generators, notably on the vectors of a basis $\underline{u}$. Denote by

$$\begin{aligned} f(u_1) &= a_{11} \cdot v_1 + a_{21} \cdot v_2 + \dots + a_{m1} v_m \\ f(u_2) &= a_{12} \cdot v_1 + a_{22} \cdot v_2 + \dots + a_{m2} v_m \\ &\ \ \vdots \\ f(u_n) &= a_{1n} \cdot v_1 + a_{2n} \cdot v_2 + \dots + a_{mn} v_m, \end{aligned}$$

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.

Matrix $A = (a_{ij})$ is called *matrix of the mapping f* in bases $\underline{u}$, $\underline{v}$.

For a general vector $u = x_1 u_1 + \dots + x_n u_n \in V$ we calculate (recall that vector addition is commutative and distributive with respect to scalar multiplication)

$$\begin{aligned} f(u) &= x_1 f(u_1) + \dots + x_n f(u_n) \\ &= x_1(a_{11}v_1 + \dots + a_{m1}v_m) + \dots + x_n(a_{1n}v_1 + \dots + a_{mn}v_m) \\ &= (x_1 a_{11} + \dots + x_n a_{1n})v_1 + \dots + (x_1 a_{m1} + \dots + x_n a_{mn})v_m. \end{aligned}$$

Using matrix multiplication we can now very easily and clearly write down the values of the mapping $f_{\underline{u},\underline{v}}(w)$ defined uniquely by the previous diagram. Recall that vector in $\mathbb{K}^r$ are understood as columns, that is, matrices of the type $r/1$

$$f_{\underline{u},\underline{v}}(\underline{u}(w)) = \underline{v}(f(w)) = A \cdot \underline{u}(w).$$

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 $f : V \to W$. If we have chosen bases of spaces V and W, every choice of a matrix of the type m/n correspond to a unique linear mapping $V \to W$ and we have shown a bijection between matrices of the corresponding dimension and linear mappings $V \to W$.

2.61. 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 some maximal set of linearly independent vectors u_i (that is, by choosing a basis).

Solution. We write the vectors u_i into the columns of a matrix and transform it using elementary row transformations. This way we obtain

$$\begin{pmatrix} -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \\ 1 & 2 & 0 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 4 & -4 & 5 \\ 0 & -7 & 7 & -7 \\ 0 & 3 & -3 & 3 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

>From that it follows that linearly independent are exactly the vectors u_1, u_2, u_4, that is, exactly the vectors corresponding to the columns which contain first non-zero number of some row. Furthermore we have (see the third column)

$$2 \cdot (-1, 3, -2, 1) - (2, -1, -1, 2) = (-4, 7, -3, 0).$$

□

2.62. In the vector space $\mathbb{R}^4$ we are given three-dimensional subspaces

$$U = \langle u_1, u_2, u_3 \rangle, \qquad V = \langle v_1, v_2, v_3 \rangle,$$

while

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \ u_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \ u_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad v_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \ v_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix},$$

$v_3 = (1, -1, -1, 1)^T$. Determine the dimension and give 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. We thus 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

$$\begin{array}{rcrcrcrcrcrcr} x_1 & + & x_2 & + & x_3 & = & y_1 & + & y_2 & + & y_3, \\ x_1 & + & x_2 & & & = & y_1 & - & y_2 & - & y_3, \\ x_1 & & & + & x_3 & = & -y_1 & + & y_2 & - & y_3, \\ & & x_2 & + & x_3 & = & -y_1 & - & y_2 & + & y_3. \end{array}$$

2.37. Matrix for changing the coordinates. If we choose V and W to be the same space, but with different bases, and for f pick the identity mapping, 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. If we then have the vector u as

$$u = x_1 u_1 + \cdots + x_n u_n,$$

that is, in coordinates with respect to $\underline{u}$ and plug for u_i their expression using the vectors from $\underline{v}$, we obtain the coordinate expression $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$ of the same vector in the basis $\underline{v}$. It is enough just to reorder the summands and express the individual scalars at the vectors of the basis.

In reality, we are doing exactly the same thing as in the previous paragraph for the special case of the identity mapping id_V on the vector space V. Matrix of this identity mapping is T and therefore the direct calculation must give $\bar{x} = T \cdot x$. The situation is depicted in the diagram

$$\begin{tikzcd} V \arrow[r, "\mathrm{id}_V"] \arrow[d, "\underline{u}"', "\simeq"] & V \arrow[d, "\simeq"', "\underline{v}"] \\ \mathbb{K}^n \arrow[r, dotted, "T=(\mathrm{id}_V)_{\underline{u},\underline{v}}"] & \mathbb{K}^n \end{tikzcd}$$

The resulting matrix T is called *matrix for changing the basis* from $\underline{u}$ of the vector space V to the basis $\underline{v}$ of the same space.

Directly from the definition we have:

Calculating the matrix for changing the basis

Proposition. *Matrix T for changing from the basis $\underline{u}$ to the basis $\underline{v}$ is obtained by taking coordinates of the vectors of the basis $\underline{u}$ expressed in the basis $\underline{v}$ are written as the columns of the matrix T.*

The role of the matrix for changing the basis is that if we know the coordinates x of the vector in the basis $\underline{u}$, then its coordinates in the basis $\underline{v}$ are obtained by multiplying the column x with the matrix for changing the basis (from the left). Because the inverse mapping for the identity mapping is again an identity mapping, the matrix for changing the basis is always invertible and its inverse is the matrix for changing the basis in the opposite direction, that is, from the basis $\underline{v}$ to the basis $\underline{u}$.

2.38. More coordinates. Now we show how to compose possible coordinate expressions of linear mapping. Let us 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 $g_{\underline{v},\underline{w}}$.

$$\begin{tikzcd} V \arrow[r, "f"] \arrow[d, "\underline{u}"', "\simeq"] & W \arrow[r, "g"] \arrow[d, "\simeq"', "\underline{v}"] & Z \arrow[d, "\simeq"', "\underline{w}"] \\ \mathbb{K}^n \arrow[r, dotted, "f_{\underline{u},\underline{v}}"] & \mathbb{K}^m \arrow[r, dotted, "g_{\underline{v},\underline{w}}"] & \mathbb{K}^k \end{tikzcd}$$

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 directly calculate (we write A for the matrix f and B for the matrix of g in the chosen

Using matrix notation of this homogeneous system (and preserving the order of the variables) we have

$$\begin{pmatrix}1&1&1&-1&-1&-1\\1&1&0&-1&1&1\\1&0&1&1&-1&1\\0&1&1&1&1&-1\end{pmatrix}\sim\begin{pmatrix}1&1&1&-1&-1&-1\\0&0&-1&0&2&2\\0&-1&0&2&0&2\\0&1&1&1&1&-1\end{pmatrix}$$

$$\sim\begin{pmatrix}1&1&1&-1&-1&-1\\0&1&1&1&1&-1\\0&0&-1&0&2&2\\0&0&1&3&1&1\end{pmatrix}\sim\begin{pmatrix}1&1&1&-1&-1&-1\\0&1&1&1&1&-1\\0&0&1&0&-2&-2\\0&0&0&1&1&1\end{pmatrix}$$

$$\sim\begin{pmatrix}1&1&1&0&0&0\\0&1&1&0&0&-2\\0&0&1&0&-2&-2\\0&0&0&1&1&1\end{pmatrix}\sim\begin{pmatrix}1&0&0&0&0&2\\0&1&0&0&2&0\\0&0&1&0&-2&-2\\0&0&0&1&1&1\end{pmatrix}.$$

We obtain a solution

$$x_1=-2t,\ x_2=-2s,\ x_3=2s+2t,\ y_1=-s-t,\ y_2=s,\ y_3=t,$$

$t, s \in \mathbb{R}$. 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

$$\dim U\cap V=2,\qquad U\cap V=\left\langle\begin{pmatrix}0\\-1\\1\\0\end{pmatrix},\begin{pmatrix}0\\-1\\0\\1\end{pmatrix}\right\rangle.$$

□

2.63. Give some basis of the subspace

$$U=\left\langle\begin{pmatrix}1&2\\3&4\\5&6\end{pmatrix},\begin{pmatrix}0&1\\2&3\\4&5\end{pmatrix},\begin{pmatrix}-1&0\\1&2\\3&4\end{pmatrix},\begin{pmatrix}-2&-1\\0&1\\2&3\end{pmatrix}\right\rangle$$

of the vector space of real matrices 3 × 2. Extend this basis to a basis of the whole space.

Solution. Let us remind that a basis of a subspace is a set of linearly independent vectors which generate the given subspace. Because

$$-1\cdot\begin{pmatrix}1&2\\3&4\\5&6\end{pmatrix}+2\cdot\begin{pmatrix}0&1\\2&3\\4&5\end{pmatrix}=\begin{pmatrix}-1&0\\1&2\\3&4\end{pmatrix},$$

$$-2\cdot\begin{pmatrix}1&2\\3&4\\5&6\end{pmatrix}+3\cdot\begin{pmatrix}0&1\\2&3\\4&5\end{pmatrix}=\begin{pmatrix}-2&-1\\0&1\\2&3\end{pmatrix},$$

the whole subspace U is generated just by the first two matrices. These are furthermore linearly independent (none is a multiple of another) and thus give a basis. If we want to extend it to a basis of the whole space of real matrices 3 × 2, we must find four more matrices (the

bases):

$$\begin{aligned}g_{\underline{v},\underline{w}}\circ f_{\underline{u},\underline{v}}(x)&=\underline{w}\circ g\circ\underline{v}^{-1}\circ\underline{v}\circ f\circ\underline{u}^{-1}\\&=B\cdot(A\cdot x)=(B\cdot A)\cdot x=(g\circ f)_{\underline{u},\underline{w}}(x)\end{aligned}$$

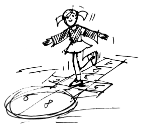
for every $x\in\mathbb{K}^n$. Composition of mappings thus corresponds to multiplication of the corresponding matrices. Note that the isomorphisms correspond exactly to invertible matrices.

The same approach gives us an answer to the question how does the matrix of the mapping change whenever we change the basis (both in the domain and in the codomain):

$$\begin{tikzcd}V \arrow[r, "\mathrm{id}_V"] \arrow[d, "\underline{u}'"', "\simeq"] & V \arrow[r, "f"] \arrow[d, "\underline{u}"', "\simeq"] & W \arrow[r, "\mathrm{id}_W"] \arrow[d, "\simeq"', "\underline{v}"] & W \arrow[d, "\simeq"', "\underline{v}'"] \\ \mathbb{K}^n \arrow[r, dotted, "T"] & \mathbb{K}^n \arrow[r, dotted, "f_{\underline{u},\underline{v}}"] & \mathbb{K}^m \arrow[r, dotted, "S^{-1}"] & \mathbb{K}^m\end{tikzcd}$$

where T is the matrix for changing the basis from $\underline{u}'$ to $\underline{u}$ and S is the matrix for changing the basis 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 linear mapping $f : V \to V$, that is, mapping that has the same space V as its domain and codomain, we express f usually with a single basis $\underline{u}$ of the space V. Then the changing of the basis to the new one $\underline{u}'$ with the matrix T for changing from $\underline{u}'$ to $\underline{u}$ the new matrix will be $A'=T^{-1}AT$.



2.39. Linear forms. A specially simple but important case of linear mappings are so-called *linear forms*. They are linear mappings from the vector space V over field of scalars $\mathbb{K}$ into the scalars $\mathbb{K}$. If we are given the coordinates on V, the assignments of a single i-th coordinate to the vectors is an example of a linear form. More precisely, for every choice of basis $\underline{v}=(v_1,\dots,v_n)$ we have at our disposal the linear forms $v_i^*: V\to\mathbb{K}$ such that $v_i^*(v_j)=\delta_{ij}$, that is, zero for distinct indices i and j and one for the same indices.

Vector space of all linear forms on V is denoted by V^* and called *dual space* of the vector space V. Let us now assume that the vector space V has finite dimension n. The basis V^* composed of assignments of individual coordinates as before is called *dual basis*. Really it is a basis of the space V^*, because these forms are clearly linearly independent (prove it!) and if α is an arbitrary form, then it holds for every vector $u=x_1v_1+\dots+x_nv_n$

$$\begin{aligned}\alpha(u)&=x_1\alpha(v_1)+\dots+x_n\alpha(v_n)\\&=\alpha(v_1)v_1^*(u)+\dots+\alpha(v_n)v_n^*(u)\end{aligned}$$

and thus the linear form α is a linear combination of the forms v_i^*.

For a fixed basis $\{1\}$ on one-dimensional space of scalars $\mathbb{K}$ are for every choice of the basis $\underline{v}$ on V the linear forms α identified with matrices of the type $1/n$, that is, with rows y. Exactly the components of these rows are coordinates of the general linear forms in the dual basis $\underline{v}^*$. Expressing such form on 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 $\underline{v}$:

$$\alpha(u)=y\cdot x=y_1x_1+\dots+y_nx_n.$$

Thus we can see that for every finitely dimensional space V is V^* isomorphic to the space V. Realisation of such isomorphism is given for instance by our choice of the dual basis for the chosen basis on the space V.

dimension of the whole space is clearly 6) such that the resulting six-tuple is linearly independent. We can use for instance the canonical basis

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

of the space of real matrices 3 × 2, which can be identified directly with $\mathbb{R}^6$. If we write down the two vectors of the basis of U and then the canonical basis of the whole space, by choosing first 6 linearly independent vectors we obtain a desired basis. If we consider for instance

$$\begin{vmatrix}1&0&0&0&0&0\\2&1&0&0&0&0\\3&2&1&0&0&0\\4&3&0&1&0&0\\5&4&0&0&1&0\\6&5&0&0&0&1\end{vmatrix}=1,$$

we can immediately add to the basis vectors

$$\begin{pmatrix}1&2\\3&4\\5&6\end{pmatrix},\quad \begin{pmatrix}0&1\\2&3\\4&5\end{pmatrix}$$

of the subspace U the matrices (vectors of the space of the matrices)

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

to a basis. Let us note that the determinant given above is easy to compute – it equals the product of all elements on the diagonal, because the matrix is in lower triangular form (everything above the diagonal is zero). □

H. Linear mappings

How to analytically describe similar mappings (for instance rotation, axial symmetry, mirror symmetry, projection of a three-dimensional space on a two-dimensional one) in the plane or in the space? How can we describe scaling of a picture? What do they have in common? They all are linear mappings. That means that they preserve certain structure of the space or a subspace. What structure? 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 some point is in some direction twice that far from the origin as some other point. We also know where do we get if we shift by some value in a given direction and then by some other value in another direction. These properties can be formalised – we speak of vectors in the plane or in the space and about their multiplication and addition. Linear mapping has the property that the image of a sum of vectors is

In this context we again meet the scalar product of a row of n scalars with a column of n scalars, as we have worked with it already in the paragraph 2.3 on the page 74.

When considering an infinitely dimensional space, things behave differently. 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 $v_i = x^i$ and as before we can define linearly independent forms v_i^*. Every formal infinite sum $\sum_{i=0}^{\infty} a_i v_i^*$ is now well-defined linear form on $\mathbb{K}[x]$, because it will be evaluated only on a finite linear combination of the basis polynomials x^i, $i = 0, 1, 2, \dots$.

The countable set of all v_i^* is thus not a basis. In reality, one can show that this dual space cannot have a countable basis.

2.40. The size of vectors and scalar product. When dealing with the geometry of the plane $\mathbb{R}^2$ in the first chapter in the paragraph 1.29 we have already worked not with just bases and linear mappings but also with the size of vectors and their angles. For defining these terms we have used the scalar product of two vectors $v = (x, y)$ and $v' = (x', y')$ in the form $u \cdot v = xx' + yy'$. Really, the actual expression for the size 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 $v = (x, y)$ and $v' = (x', y')$ is in planar geometry given by the relation

$$\cos\varphi = \frac{xx' + yy'}{\|v\|\|v'\|}.$$

Note that this scalar product is linear in every of its arguments, we denote it by $u \cdot v$ or by $\langle v, v'\rangle$. Scalar product defined in such way is symmetric in its arguments and of course that it holds that $\|v\| = 0$ if and only if $v = 0$. >From our considerations it can be seen that in the Euclidean plane two vectors are perpendicular whenever their scalar product is zero.

In the case of real vector space of any dimension we shall try a similar approach, because the concept of the angle of two vectors is clearly always two-dimensional (we want the angle to be the same in the two-dimensional space containing u and v). In the following paragraphs, we shall consider only finitely dimensional vector spaces over real scalars $\mathbb{R}$.

Scalar product and perpendicularity

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 its arguments, 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 size 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 of orthogonal vectors only is called *orthogonal basis*. If the vectors are additionally normalised, it is then *orthonormalised basis*.

a sum of the images of the vectors and the image of a multiple of a vector is the multiple of the image of the vector. These properties are shared among the mappings stated at the start of this paragraph. Such a mapping is then uniquely determined by its behaviour on vectors of a basis (in the plane by the image of two vectors not on the same line, in the space by the image of three vectors not in the same plane).

And how to write down some linear mapping f on a vector space V? Let us start for simplicity 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 arbitrary point with coordinates (u, v): $f((u, v)) = f(u(1, 0) + v(0, 1)) = uf(1, 0) + vf(1, 0) = (ua, ub) + (vc, vd) = (au + cv, bu + dv)$, which can be efficiently written down as follows:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} au + cv \\ bu + dv \end{pmatrix}$$

Linear mapping is thus a mapping uniquely determined 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 $\begin{pmatrix} ae + fc & be + df \\ ag + ch & bg + dh \end{pmatrix}$.

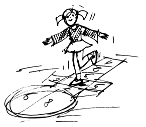
This leads us to the definition of matrix multiplication in exactly this way, that is, we want that an application of a mapping on a vector is given by the matrix multiplication of the matrix of the mapping with the given vector, and that mapping composition is given by the product of the corresponding matrices. It works analogously in the spaces of higher dimension. Further, this again shows what has already been proven in (2.5), that is, that matrix multiplication is associative but not commutative, because it is so with mapping composition. That is another of the motivation why one should investigate vector spaces.

Let us now recall that already in the first chapter we have worked with matrices of some linear mappings in the plane $\mathbb{R}^2$, notably with the rotation around a point and with axial symmetry (see 1.31 and 1.32).

Let us now try to write down matrices of linear mappings from $\mathbb{R}^3$ to $\mathbb{R}^3$. How does the matrix of a rotation in three dimensions look like? Let us begin with special (easier for description) rotations about coordinate axes:

2.64. Matrix of rotation about coordinate axes in $\mathbb{R}^3$. We write down matrices of the rotations by the angle φ, gradually about the (oriented) axes x, y and z in $\mathbb{R}^3$.

Scalar product is very often denoted by the common dot, that is, $\langle u, v\rangle = u{\cdot}v$. From the context it is then necessary to recognise whether it is a product of two vectors (result is a scalar then) or something different (we have denoted the product the matrices and the product of scalars in the same way sometimes).



Because scalar product is linear in each of its arguments, it is completely determined by its values on tuples of basis vectors. Really, let us choose a basis $\underline{u} = (u_1, \dots, u_n)$ of the space V and denote

$$s_{ij} = \langle u_i, u_j\rangle.$$

Then from the symmetry of scalar product we have $s_{ij} = s_{ji}$ and from the linearity of the product in each of its arguments we get

$$\Big\langle \sum_i x_i u_i, \sum_j y_j u_j \Big\rangle = \sum_{i,j} x_i y_j \langle u_i, u_j\rangle = \sum_{i,j} s_{ij} x_i y_j.$$

If the basis is orthonormal, the matrix S is the unit matrix. This proves the following useful claim:

SCALAR PRODUCT AND ORTHONORMAL BASIS

Proposition. *Scalar product is in every orthonormal basis given in coordinates by the expression*

$$\langle x, y\rangle = x^T \cdot y.$$

For every general basis of the space V there is symmetric matrix S such that the coordinate expression of the scalar product is

$$\langle x, y\rangle = x^T \cdot S \cdot y.$$

2.41. Orthogonal complements and projections. For every fixed subspace $W \subset V$ in a space with scalar product we define its *orthogonal complement* as follows

$$W^\perp = \{u \in V;\ u \perp v \text{ for all } v \in W\}.$$

Directly from the definition it is clear that $W^\perp$ is vector subspace. If $W \subset V$ has basis $(u_1, \dots, u_k)$ the condition 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 $\langle u, u\rangle = 0$ and thus also $u = 0$ due to the definition of scalar product. Clearly then the whole space V is the direct sum

$$V = W \oplus W^\perp.$$

Linear mapping $f : V \to V$ on any vector space is called *projection*, if we have

$$f \circ f = f.$$

In such case 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 $v = 0$. The previous sum of subspaces is then direct. We say that f is a projection on 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 component in W and in U and forget the second one.

If V has a scalar product, we say that the projection is perpendicular if the kernel is perpendicular to the image.

Solution. When rotating any particular point about the given axis (say x), the corresponding coordinate (x) does not change and the remaining two coordinates are then given by the rotation in the plane which we already know (a matrix of the type $2/times2$).

Thus we gradually 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 y:

$$\begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix}$$

rotation about the axis x:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}.$$

The sign at φ in the matrix for rotation about y is different. 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 so-called "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 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 through all three axes. □

The knowledge of matrices allows us to write the matrix of rotation about any (oriented) axis. Let us start with a specific example:

2.65. Find the matrix of the rotation in the positive sense through 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 can be 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).

Every subspace $W \neq V$ thus defines an *perpendicular projection* on W. It is a projection on 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 $u_W + u_{W^\perp}$ on u_W.

2.42. Existence of orthonormal basis. Note that on every finitely dimensional real vector space there definitely exist scalar products. Just pick any basis, call it orthonormal and we immediately have a scalar product. In this basis the scalar products are computed as in the formula in the Theorem 2.40.

But we can do it in other way too. If we are given scalar product on a vector space V, we can easily use some suitable perpendicular projections and transform any basis into orthonormal one. It is called *Gramm-Schmidt orthogonalisation process*. The point of this procedure is to transform a given sequence of nonzero generators $v_1, \dots, v_k$ of a finitely dimensional space V into an orthogonal set of nonzero generators for V.

Gramm-Schmidt orthogonalisation

Proposition. *Let $(u_1, \dots, u_k)$ be a linearly independent k-tuple of vectors of a space V with scalar product. Then there exists an orthogonal system of vectors $(v_1, \dots, v_k)$ such that $v_i \in \langle u_1, \dots, u_i\rangle$, $i = 1, \dots, k$. We obtain it by the following procedure:*

- *The independence of the vectors u_i ensures that $u_1 \neq 0$; we choose $v_1 = u_1$.*
- *If we have already constructed the vectors $v_1, \dots, v_\ell$ of required properties, we choose $v_{\ell+1} = u_{\ell+1} + a_1 v_1 + \dots + a_\ell v_\ell$, where $a_i = -\frac{\langle u_{\ell+1}, v_i\rangle}{\|v_i\|^2}$.*

Proof. Let us begin with the first (nonzero) vector v_1 and calculate the perpendicular projection v_2 on do

$$\langle v_1\rangle^\perp \subset \langle\{v_1, v_2\}\rangle.$$

The result is nonzero if and only if v_2 is independent on v_1. In all further steps we work similarly.

In the ℓ-th we want that for $v_{\ell+1} = u_{\ell+1} + a_1 v_1 + \dots + a_\ell v_\ell$ holds $\langle v_{\ell+1}, v_i\rangle = 0$ for all $i = 1, \dots, \ell$. That implies

$$0 = \langle u_{\ell+1} + a_1 v_1 + \dots + a_\ell v_\ell, v_i\rangle = \langle u_{\ell+1}, v_i\rangle + a_i\langle v_i, v_i\rangle$$

and we can see that the vectors with desired properties are determined uniquely up to a scalar multiple. □

Whenever we have an orthogonal basis of a vector space V, we just have to normalise the vectors in order to obtain an orthonormal basis. Thus we have proven:

Corollary. *On every finitely dimensional real vector space with scalar product there exist an orthonormal basis.*

In orthonormal basis the coordinates and perpendicular projections are very easy to calculate. Really, let us have an orthonormal basis $(e_1, \dots, e_n)$ of a space V. Then every vector $v = x_1 e_1 + \dots + x_n e_n$ satisfies

$$\langle e_i, v\rangle = \langle e_i, x_1 e_1 + \dots + x_n e_n\rangle = x_i$$

and it always holds that

$$v = \langle e_1, v\rangle e_1 + \dots + \langle e_n, v\rangle e_n. \tag{2.3}$$

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

$$\begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} =$$
$$= \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{6}}{4} \\ \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{6}}{4} \\ -\frac{\sqrt{6}}{4} & \frac{\sqrt{6}}{4} & \frac{1}{2} \end{pmatrix}$$

□

2.66. 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 arbitrary unit vector (x, y, z). Rotation in the positive sense through 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 $(0, y, z)$. Matrix of this rotation is

$$R_1 = \begin{pmatrix} x/\sqrt{1-z^2} & y/\sqrt{1-z^2} & 0 \\ -y/\sqrt{1-z^2} & x/\sqrt{1-z^2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

If we are given a subspace $W \subset V$ and its orthonormal basis $(e_1, \dots, e_k)$, we can surely extended it to an orthonormal basis $(e_1, \dots, e_n)$ of the whole V. Perpendicular projection of a general vector $v \in V$ on W is then given by the relation

$$v \mapsto \langle e_1, v\rangle e_1 + \dots + \langle e_n, v\rangle e_k.$$

For perpendicular projection it is enough to know just the orthonormal basis of the subspace W, on which we are projecting.

Let us also note than in general projections f on a subspace W along U and projections g on U along W tied with the relation $g = \mathrm{id}_V - f$. When dealing with perpendicular projections on a given subspace W, it is always more efficient to calculate the orthonormal basis of the space which has smaller dimension (that is, for either W or $W^\perp$).

Let us also note that the existence of an orthonormal basis ensures that for every real space V of dimension n with scalar product there exists a linear mapping which is an isomorphism between V and the space $\mathbb{R}^n$ with standard scalar product. Similarly it has been shown already in the Theorem 2.40, where we have shown that the desired isomorphism is exactly the coordinate assignment. In words – in orthonormal basis the scalar product with coordinates is computed by the same formula as the standard scalar product in $\mathbb{R}^n$.

We shall return to the questions of the size of a vector and to projections in the following chapter in more general context.

2.43. Angle of two vectors. As we have already noted, the angle of 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 and if we choose 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. Even without choosing the basis it must hold that:

ANGLE OF TWO VECTORS

Angle φ of two vectors v and w in a vector space with scalar product is given by the relation

$$\cos\varphi = \frac{\langle v, w\rangle}{\|v\|\|w\|}.$$

Angle defined in this way does not depend on the order of the vectors v, w and is in the interval $0 \le \varphi \le \pi$.

We shall return to the scalar products and angles between vectors in further chapters.

2.44. Multilinear forms. Scalar product was given as a mapping from the product of two copies of a vector space V into the space of scalars, which was linear in each of its arguments. Similarly, we will work with mappings from the product of k copies of a vector space V into the scalars, which are linear in each of its k arguments. We speak of k-*linear* forms.

Most often we will meet *bilinear forms*, that is, the case $\alpha : V \times V \to \mathbb{K}$, where for any four vectors u, v, w, z and scalars a,

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 $(1, 0, 0)$. 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

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix},$$

iv) rotation $\mathcal{R}_2^{-1}$ with the matrix R_2^{-1},

v) rotation $\mathcal{R}_1^{-1}$ with the matrix R_1^{-1}.

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:

$R_1^{-1} \cdot R_2^{-1} \cdot R_3 \cdot R_2 \cdot R_1 =$

$$\begin{pmatrix} \cos\varphi + (1-\cos\varphi)x^2 & (1-\cos\varphi)xy - z\sin\varphi & (1-\cos\varphi)xz + y\sin\varphi \\ yx(1-\cos\varphi) + z\sin\varphi & \cos\varphi + (1-\cos\varphi)y^2 & (1-\cos\varphi)yz - x\sin\varphi \\ zx(1-\cos\varphi) - y\sin\varphi & (1-\cos\varphi)zy + x\sin\varphi & \cos\varphi + (1-\cos\varphi)z^2 \end{pmatrix}$$

□

2.67. We are given a linear mapping $\mathbb{R}^3 \to \mathbb{R}^3$ 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 under the basis
$(f_1, f_2, f_3) = ((1, 1, 0), (-1, 1, 1), (2, 0, 1))$.

Solution. The transition matrix T for changing the basis from the basis $\underline{f} = (f_1, f_2, f_3)$ to the standard basis, that is, to the basis given by the vectors $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, can be obtained, according to the Claim 2.25, 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

$$T = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

b, c and d it holds, as with the ordinary scalar product, that

$$\begin{aligned} \alpha(au + bv, cw + dz) &= ac\,\alpha(u, w) + ad\,\alpha(u, z) \\ &\quad + bc\,\alpha(v, w) + bd\,\alpha(v, z). \end{aligned}$$

If we additionally have

$$\alpha(u, w) = \alpha(w, u),$$

we speak of *symmetric bilinear form*. If exchange of the arguments leads to opposite sign of the result, we speak of *antisymmetric bilinear form*.

Already in the planar geometry we have defined determinant as a bilinear antisymmetric form α, that is, $\alpha(u, w) = -\alpha(w, u)$. In general we know thanks to the theorem 2.17 that determinant in the dimension n can be seen as 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 on an example with $k = 2$, where it will really correspond to matrices (as we have defined them).

Matrix of bilinear form

If we choose basis $\underline{u}$ on V and define for a given bilinear form α scalars $a_{ij} = \alpha(u_i, u_j)$ then we obviously obtain for vectors v, w with coordinates x and y (as columns of coordinates)

$$\alpha(v, w) = \sum_{i,j=1}^{n} a_{ij} x_i y_j = y^T \cdot A \cdot x,$$

where A is a matrix $A = (a_{ij})$.

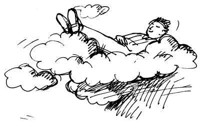
Directly from the definition of the matrix of bilinear form we see that the form is symmetric or antisymmetric if and only if the corresponding matrix has this property.

Every bilinear form α on vector space V defines a mapping $V \to V^*$, $v \mapsto \alpha(\ , v)$, that is, plugging a fixed vector in the second argument we obtain a linear form which is the image of this vector. If we choose a fixed basis on a finitely dimensional space V and a dual basis V^*, then we have a mapping

$$y \mapsto (x \mapsto y^T \cdot A \cdot x).$$

4. Properties of linear mappings

More detailed analysis of properties of types of linear mappings will now lead us to a better understanding of tools which vector spaces give us for modelling of linear processes and systems.



2.45. Let us begin with four examples in the lowest interesting dimension. In the standard basis of the plane $\mathbb{R}^2$ with the standard scalar product we consider the following matrices of mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, C = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

The matrix A gives a perpendicular projection along the subspace

$$W \subset \{(0, a);\ a \in \mathbb{R}\} \subset \mathbb{R}^2$$

Transition matrix for changing the basis from the standard basis to the basis $\underline{f}$ is then given by

$$T^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{1}{2} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

Matrix of the mapping in the basis $\underline{f}$ is then given by

$$T^{-1}AT = \begin{pmatrix} \frac{1}{4} & 2 & -\frac{3}{4} \\ \frac{5}{4} & 0 & \frac{7}{4} \\ \frac{3}{4} & -2 & \frac{9}{4} \end{pmatrix}.$$

□

2.68. 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 $1 + x^2, x, x + x^2$.

Solution. $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 2 & 1 & 3 \\ 0 & -1 & -1 \end{pmatrix}$. □

2.69. 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, give 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 shall note that the matrix of the rotation in the positive sense through 90° in some orthonormal basis in $\mathbb{R}^2$ is $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. In orthogonal basis with sizes of the vectors k, l 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 sizes $\sqrt{2}$ 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

on the subspace

$$V \subset \{(a, 0);\ a \in \mathbb{R}\} \subset \mathbb{R}^2,$$

that is, the projection on the x-axis along the y-axis Evidently for this mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ it holds that $f \circ f = f$ and thus the restriction $f|_V$ of the given mapping on its codomain is 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 it as a mapping of differentiation of polynomials $\mathbb{R}_1[x]$ of degree at most one in the basis $(1, x)$ (with differentiation we shall deal in the chapter five, see **??**).

The matrix C gives a mapping f, which enlarges the first vector of the basis a-times, the second 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 (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 a matrix of rotation through the right angle in the standard basis and on the first sight we can see that none of the one-dimensional subspaces is not preserved under this mapping.

Such rotation is a bijection of the plane to 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 on both the domain and the codomain.

Let us see the matrix D as a matrix of the mapping $g : \mathbb{C}^2 \to \mathbb{C}^2$ in 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

$$g(u) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i \cdot u,$$

$$g(v) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} -1 \\ -i \end{pmatrix} = -i \cdot v.$$

That means that in the basis (u, v) on $\mathbb{C}^2$ the mapping g has the matrix

$$K = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

and note that the this complex analogy to the case of matrix C has on the diagonal the 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 this number in polar form gives 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 complex conjugate of u. We are interested in the restriction of the mapping g on the real vector space $V = \mathbb{R}^2 \cap \langle u, v \rangle \subset \mathbb{C}^2$. Evidently is

$$V = \langle u + \bar{u}, i(u - \bar{u}) \rangle = \langle x_u, -y_u \rangle$$

$$R = T \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -\sqrt{3} \\ 0 & 1/\sqrt{3} & 0 \end{pmatrix} \cdot T^{-1}$$
$$= \begin{pmatrix} 1/3 & 1/3-\sqrt{3}/3 & 1/3+\sqrt{3}/3 \\ 1/3+\sqrt{3}/3 & 1/3 & 1/3-\sqrt{3}/3 \\ 1/3-\sqrt{3}/3 & 1/3+\sqrt{3}/3 & 1/3 \end{pmatrix}$$

This result can be checked by plugging into the matrix of general rotation (||2.66||). By normalising the vector $(1, 1, 1)$ we obtain the vector $(x, y, z) = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, $\cos(\varphi) = 0$, $\sin(\varphi) = 1$. □

2.70. In $\mathbb{R}^3$ determine the matrix of rotation through the angle 120° in the positive sense about the vector $(1, 0, 1)$ (it is enough to give in the form of matrix product). ○

2.71. Matrix of general rotation revisited. Let us try to derive the matrix of (general) rotation from (||2.66||) through the angle φ in the positive sense about the unit vector (x, y, z) in a different way, analogically to the previous exercise. In the basis $\underline{f} = ((x, y, z), (-y, x, 0), (zx, zy, z^2-1))$, 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 $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ \cos(\varphi) & \sin(\varphi) & 0 \end{pmatrix}$. 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

$$R = T \cdot R \cdot T^{-1} =$$
$$\begin{pmatrix} \cos\varphi + (1-\cos\varphi)x^2 & (1-\cos\varphi)xy - z\sin\varphi & (1-\cos\varphi)xz + y\sin\varphi \\ yx(1-\cos\varphi) + z\sin\varphi & \cos\varphi + (1-\cos\varphi)y^2 & (1-\cos\varphi)yz - x\sin\varphi \\ zx(1-\cos\varphi) - y\sin\varphi & (1-\cos\varphi)zy + x\sin\varphi & \cos\varphi + (1-\cos\varphi)z^2 \end{pmatrix}$$

When doing multiplication and simplification 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 modelling of processes and systems.

the whole plane $\mathbb{R}^2$. The restriction of g on this plane is exactly the original mapping given by the matrix A and from the definition of multiplication by the complex unit it is a 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, and realise also why exchanging the order of the vectors u and v leads to the same result, although in a different real basis!).

2.46. Eigenvalues and eigenvectors of mappings. Key to the description of mappings in the previous examples were answers to the question "what are the vectors satisfying the equation $f(u) = a \cdot u$ for some suitable scalars a?".

Let us fix a linear mapping $f : V \to V$ on a vector space of dimension n over scalars $\mathbb{K}$. If we imagine such equality written in coordinates, that is, using the matrix of the mapping A in some bases, it is an expression

$$A \cdot x - a \cdot x = (A - a \cdot E) \cdot x = 0.$$

>From the previous we know that such a system of equations has the only 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 (see Theorem 2.23)

$$\det(A - a \cdot E) = 0. \qquad (2.4)$$

If we consider $\lambda = a$ a variable in the previous scalar equation, we are actually looking for roots of polynomial of n-th degree. As we have seen in the case of the matrix D, the roots may exist, but do not have to depending to the field of scalars $\mathbb{K}$ we are having.

Eigenvalues and eigenvectors

Scalars λ satisfying the equation $f(u) = \lambda \cdot u$ for a nonzero vector $u \in V$ are called *eigenvalues of mapping* f, the corresponding nonzero vectors u then *eigenvectors of mapping* f.

If u, v are eigenvectors associated with the same eigenvalue λ, then for every linear combination of u and v it holds

$$f(au + bv) = af(u) + bf(v) = \lambda(au + bv).$$

Therefore the eigenvectors associated with the same eigenvalue λ form along with a zero vector a nontrivial vector subspace V_λ, that is, eigenspace associated with λ. For instance, if $\lambda = 0$ is an eigenvalue, the kernel Ker f is a eigenspace V_0.

>From the definition of the eigenvalues it is clear that their computation cannot depend on the choice of the basis and the matrix of the mapping f. Indeed, as a direct corollary of the transformation properties from the paragraph 2.38 and Cauchy theorem 2.19 for calculation of the determinant of product we obtain by choosing different coordinates a matrix $A' = P^{-1}AP$ with invertible matrix P and

$$\begin{aligned} |P^{-1}AP - \lambda E| &= |P^{-1}AP - P^{-1}\lambda EP| \\ &= |P^{-1}(A - \lambda E)P| = |P^{-1}||(A - \lambda E||P| \\ &= |A - \lambda E|, \end{aligned}$$

because scalar multiplication is commutative and $|P^{-1}| = |P|^{-1}$.

>From these reason we use for matrices and mappings the same terminology:

2.72. 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 swaps basis vectors, that is, $(1, 0) \mapsto (0, 1)$ and $(0, 1) \mapsto (1, 0)$ and the matrix of conjugation under this basis is $\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$.

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: $\begin{pmatrix}2 & -1\\ 1 & 2\end{pmatrix}$.

Now let us 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 have obtained 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? □

2.73. 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. Let us first note that the given subspace is a plane going through the origin with the normal vector $u_3 = (1, 1, 1)$. The ordered triple $(1, 1, 1)$ is clearly a solution to the system

$$\begin{aligned} -x_1 + x_2 \phantom{{}+x_3} &= 0,\\ -x_1 \phantom{{}+x_2} + x_3 &= 0, \end{aligned}$$

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

Characteristic polynomial of matrix and mapping

For a matrix A of dimension n over $\mathbb{K}$ we call the polynomial $|A - \lambda E| \in \mathbb{K}_n[\lambda]$ *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, its coefficients at individual powers of the variable λ are scalars expressing the properties of f, that is, they cannot depend on the choice of the basis. Notably as a simple exercise for calculating determinants we express the coefficients at the highest and lowest powers (we assume $\dim V = n$ and the matrix of the mapping $A = (a_{ij})$ to be in a certain basis):

$$\begin{aligned} |A - \lambda \cdot E| = (-1)^n\lambda^n + (-1)^{n-1}(a_{11} + \dots + a_{nn}) \cdot \lambda^{n-1}\\ + \dots + |A| \cdot \lambda^0. \end{aligned}$$

Coefficient at the highest power says only whether the dimension of the space V is even or odd. We have already noted that the determinant of the matrix of a mapping expresses how the given linear mapping scales the volume.

Interesting is that the sum of the diagonal elements of the matrix of a mapping does not depend on the choice of basis. We call it the *trace of matrix* and denote it by TrA. *Trace of mapping* is defined as a trace of the matrix in an arbitrary basis. In reality this is not so surprising, because in the eight chapter we show an example to illustrate a method of differential calculus, which shows that the trace is actually a linear approximation of the determinant in the neighbourhood of the unit matrix, see **??**.

In the following we show a few important properties of eigenspaces.

2.47. Theorem. *Eigenvectors of linear mappings $f : V \to V$ associated to different eigenvectors are linearly independent.*

Proof. Let $a_1, \dots, a_k$ be distinct eigenvalues of the mapping f and $u_1, \dots, u_k$ eigenvectors with these eigenvalues. We do the proof by induction on the number of linearly independent vectors among the chosen ones. Assume that $u_1, \dots, u_\ell$ are linearly independent and $u_{l+1} = \sum_i c_i u_i$ is their linear combinations. At least $\ell = 1$ can be chosen, 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 eigenvalues are nonzero and at least one coefficient c_i is nonzero. That is a contradiction with the assumed linear independence $u_1, \dots, u_\ell$, therefore also the vector u_{l+1} must be linearly independent of the others. □

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

$$A = \begin{pmatrix}-1 & -1 & 1\\ 1 & 0 & 1\\ 0 & 1 & 1\end{pmatrix} \cdot \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{pmatrix} \cdot \begin{pmatrix}-\frac{1}{3} & \frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3}\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{pmatrix}$$

$$= \begin{pmatrix}\frac{2}{3} & -\frac{1}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3}\end{pmatrix}.$$

□

2.74. In the vector space $\mathbb{R}^3$ determine the matrix of the orthogonal projection onto the plane $x + y - 2z = 0$. ○

2.75. In the vector space $\mathbb{R}^3$ determine the matrix of the orthogonal projection on the plane $2x - y + 2z = 0$. ○

I. Bases and inner products

Using the inner product we can solve in a different (better?) way problems we were able to solve already using changes of coordinates.

2.76. 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 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 2.3) as

$$\frac{(\mathbf{x}, (1,1,1))}{|(1,1,1)|^2} = (\frac{x_1 + x_2 + x_3}{3}, \frac{x_1 + x_2 + x_3}{3}, \frac{x_1 + x_2 + x_3}{3}).$$

The resulting mapping is thus

$$\mathbf{x} - \mathbf{p} = (\frac{2x_1}{3} - \frac{x_2 + x_3}{3}, \frac{2x_2}{3} - \frac{x_1 + x_3}{3}, \frac{2x_3}{3} - \frac{x_1 + x_2}{3}) =$$

$$= \begin{pmatrix}\frac{2}{3} & -\frac{1}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3}\end{pmatrix} \begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}.$$

We have (correctly) obtained the same matrix as in the exercise ‖2.73‖.

□

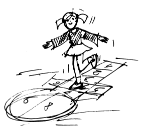
The just proved theorem can be seen as a decomposition of a linear mapping f into a sum of simple mappings. For distinct eigenvalues λ_i of the characteristic polynomial we obtain one-dimensional eigenspaces V_{λ_i}. Each of them then describes a projection on this invariant one-dimensional subspace, where the mapping is given just as multiplication by the eigenvalue λ_i. The whole subspace V is then decomposed into a direct sum of individual eigenspaces. Furthermore, this decomposition can be easily calculated:

Basis of eigenvectors

Corollary. *If there exists n mutually distinct roots λ_i of the characteristic polynomial of the mapping $f : V \to V$ on n-dimensional space V, then there exists a decomposition of V into a direct sum of eigenspaces of dimension 1. That means that there exists a basis of V composed only of eigenvectors and in this basis f has diagonal matrix. This basis is uniquely determined up to the order of the elements.*

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 a matrix of f in a chosen basis.

2.48. Invariant subspaces. We have seen that every eigenvector v of the mapping $f : V \to V$ generates a subspace $\langle v \rangle \subset V$, which is preserved by the mapping f.



In more generality, we say that a vector subspace $W \subset V$ is *invariant subspace* for a linear mapping f, if it holds that $f(W) \subset W$.

If V is a finitely dimensional vector space and we choose some basis $(u_1, \dots, u_k)$ of a subspace W, we can always extend it to a basis $(u_1, \dots, u_k, u_{k+1}, \dots, u_n)$ of the whole space V and in every such basis has our mapping the matrix A of the form

$$A = \begin{pmatrix}B & C\\ 0 & D\end{pmatrix} \tag{2.5}$$

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 in some basis $(u_1, \dots, u_n)$ the matrix of the mapping f is of the form (2.5), $W = \langle u_1, \dots, u_k \rangle$ is an invariant subspace of the mapping f.

Of course that in our matrix of the mapping (2.5) a submatrix C is zero if and only if the subspace $\langle u_{k+1}, \dots, u_n \rangle$ generated by the added vectors of the basis invariant.

>From this point of view the eigenspaces of the mapping are extremal case of invariant subspaces and notably in the case of existence of $n = \dim V$ distinct eigenvalues of the mapping f we obtain a decomposition of V into direct sum of n eigenspaces. In a suitable basis formed of the eigenvectors the mapping has then diagonal form with eigenvalues on the diagonal.

2.49. Orthogonal mappings. Let us now have a look on the special case of the mapping $f : V \to W$ between spaces with scalar products, which preserve sizes for all vectors $u \in V$.

2.77. In $\mathbb{R}^3$ a standard coordinate system is considered. 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. At what point does he think the candle is? ○

Solution. $[4, 3, -5]$ □

2.78. Find the matrix of the reflection with respect to the plane $x + y + z = 0$.

Solution. >From the equation of the plane we determine its unit normal vector. In our case it is $n = \frac{1}{\sqrt{3}}(1, 1, 1)$. The reflection Z of the vector v can then be expressed by $Zv = v - 2(v.n)n = (1 - 2nn^T)v$ (for the standard scalar product we have $v.n = vn^T$). The matrix of the reflection is then

$$1 - 2nn^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 2 \cdot \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{pmatrix}$$

□

Using the inner product we can determine the (angular) deflection of the vectors:

2.79. 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. □

2.80. Determine the cosine of the deflection of the lines p, q in $\mathbb{R}^3$ given by the equations

$$\begin{aligned} p \quad : \quad & -2x + y + z = 1 \\ & x + 3y - 4z = 5 \\ q \quad : \quad & x - y = -2 \\ & z = 6 \end{aligned}$$

2.81. Let a line be given:

$$p : [1, 1] + (4, 1)t, \ t \in \mathbb{R}$$

Determine the parametrical expression of all lines q that pass through the origin and have deflection 60° with the line p. ○

Definition of orthogonal mapping

Linear mapping $f : V \to W$ between spaces with scalar product is called *orthogonal mapping*, if for all $u \in V$

$$\langle f(u), f(u) \rangle = \langle u, u \rangle.$$

>From the linearity of f and from the symmetry of the scalar product follows that for all tuples of vectors the following equality holds:

$$\begin{aligned} \langle f(u + v), f(u + v) \rangle = \langle f(u), f(u) \rangle + \langle f(v), f(v) \rangle \\ + 2\langle f(u), f(v) \rangle. \end{aligned}$$

Therefore all orthogonal mappings satisfy also seemingly stronger condition that for all vectors $u, v \in V$ it holds that

$$\langle f(u), f(v) \rangle = \langle u, v \rangle.$$

In the initial discussion about the geometry in the plane we have proved in the Theorem 1.33 that a linear mapping $\mathbb{R}^2 \to \mathbb{R}^2$ preserves sizes 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 mapping $f : V \to W$ must be always injective, because the condition $\langle f(u), f(u) \rangle = 0$ means also $\langle u, u \rangle = 0$ and thus $u = 0$. In such case is then the dimension of the range always at least as big as the dimension of the domain of f. But then the both dimensions equal and we know that $f : V \to \operatorname{Im} f$ is a bijection. If $\operatorname{Im} f \neq W$, we extend the orthonormal basis of the image of f to an orthonormal basis of the target space and the matrix of the mapping then contains a square regular submatrix A along with zero rows so that it has the required size. Without loss of generality we can assume that $W = V$.

Our condition for the matrix of orthogonal mapping says in orthonormal basis that for all vectors for all vectors x and y in the space $\mathbb{K}^n$ the following:

$$(A \cdot x)^T \cdot (A \cdot y) = x^T \cdot (A^T \cdot A) \cdot y = x^T \cdot y.$$

By specially choosing x and y to be the vectors of the standard basis we directly obtain that $A^T \cdot A = E$, that is, the same result as in the dimension two. Thus we have obtained the following theorem:

Matrix of orthogonal mappings

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) it has the matrix A satisfying $A^T = A^{-1}$.*

Proof. Indeed, if f preserves sizes, it must have the listed property in every orthonormal basis. On the other hand, the previous calculations show that this property for matrix in one basis ensures size preservation. □

Square matrices which satisfy the equality $A^T = A^{-1}$ are called *orthogonal matrices*.

Corollary of the previous theorem is also a description of all matrices S of basis changing. Each must give a mapping $\mathbb{K}^n \to \mathbb{K}^n$ that preserves sizes and thus satisfies the condition $S^{-1} = S^T$. When changing from one orthonormal basis to another the matrix of any linear mapping changes according to the relation

2.82. Using the Gram-Schmidt orthogonalisation obtain the orthogonal basis of the subspace

$$U = \{(x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4;\ x_1 + x_2 + x_3 + x_4 = 0\}$$

of the space $\mathbb{R}^4$.

Solution. The set of solutions of the given homogeneous linear equation is clearly a vector space with the basis

$$u_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Vectors of the orthogonal basis obtained using the Gram-Schmidt orthogonalisation process shall be denoted v_1, v_2, v_3. Let us first set $v_1 = u_1$. Further, 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, let us choose a multiple $v_2 = (-1, -1, 2, 0)^T$. Further, 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

$$v_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 3 \end{pmatrix}.$$

Let us add that 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

$$(-1, 1, 1, -1)^T, \quad (1, -1, 1, -1)^T, \quad (-1, -1, 1, 1)^T.$$

□

2.83. Write down some basis of the real vector space of the matrices 3×3 over $\mathbb{R}$ with zero trace (the sum of the elements on the diagonal) and write the coordinates of the matrix

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix}$$

in this basis.

2.84. Define some 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. ○

$$A' = S^T A S.$$

2.50. Decomposition of an orthogonal mapping. Let us now have a more detailed look on eigenvectors and eigenvalues of orthogonal mappings on a real vector space V with scalar product.

Consider fixed orthogonal mapping $f : V \to V$ with matrix A in some orthonormal basis and let us try to continue as with the matrix D of rotation, as in the example 2.45.

But let us first have a general look on invariant subspaces of orthogonal mappings and their orthogonal complements. If for any subspace $W \subset V$ and orthogonal mapping $f : V \to V$ it holds that $f(W) \subset W$, then also for all $v \in W^\perp$ it holds that $w \in W$

$$\langle f(v), w\rangle = \langle f(v), f \circ f^{-1}(w)\rangle = \langle v, f^{-1}(w)\rangle = 0$$

because also $f^{-1}(w) \in W$. But that means that also $f(W^\perp) \subset W^\perp$. We have thus proved a simple but important proposition:

Proposition. *Orthogonal complement of invariant subspace is also invariant.*

If eigenvalues of orthogonal mapping are real, this claim ensures that there always exists a basis V of eigenvectors. Indeed, restriction of f on the orthogonal complement of an invariant subspace is again an orthogonal mapping, therefore we can put into the basis one eigenvector after another, until we obtain the whole decomposition of V. However, mostly the eigenvalues of orthogonal mappings are not real. We again need to make a trip into complex vector spaces. Let us formulate a result right away:

Orthogonal mapping decomposition

Theorem. *Let $f : V \to V$ be an orthogonal mapping on a vector space V with scalar product. Then all the roots of the characteristic polynomial f have size one and there exists a decomposition of V into one-dimensional eigenspaces corresponding to the eigenvalues $\lambda = \pm 1$ and two-dimensional subspaces $P_{\lambda,\bar{\lambda}}$, where f acts by rotating through 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 orthogonal matrix A which we can see as a matrix of a linear mapping on a complex space $\mathbb{C}^m$ (which just happens to have all coefficients real). Necessary there exist exactly m (complex) roots of the characteristic polynomial, counting their algebraic multiplicity (see the Fundamental theorem of algebra, **??**). Furthermore, because the characteristic polynomial of the mapping has only real coefficients, the roots are either real or there is a tuple of roots which are complex conjugates λ and $\bar{\lambda}$. The associated eigenvectors in $\mathbb{C}^m$ for such tuple of complex conjugates are actually a solutions to two systems of linear homogeneous equations which are also complex conjugates of each other – the corresponding matrices of the systems are all real except for the eigenvalues. Therefore also the solutions of this systems are complex conjugates.

Now we use the fact that for every invariant subspace its orthogonal complement is also invariant. We first find the

2.85. Determine some basis of 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 product express the size of the matrix

$$\begin{pmatrix} 0 & 3 & 1 & 0 \\ -3 & 0 & 1 & 2 \\ -1 & -1 & 0 & 2 \\ 0 & -2 & -2 & 0 \end{pmatrix}$$

.

2.86. 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. Orthogonal complement $U^\perp$ consists of exactly those vectors that are perpendicular to every solution of the system

$$\begin{array}{ccccccc} x_1 & & - & x_3 & & = & 0, \\ & x_2 & - & x_3 & - \ 6x_4 & = & 0. \end{array}$$

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

$$U^\perp = \{a \cdot (1, 0, -1, 0) + b \cdot (0, 1, -1, -6);\ a, b \in \mathbb{R}\}.$$

□

2.87. Determine whether the subspaces $U = \langle(2, 1, 2, 2)\rangle$ a $V = \langle(-1, 0, -1, 2)\,,\ (-1, 0, 1, 0)\,,\ (0, 0, 1, -1)\rangle$ of the space $\mathbb{R}^4$ are orthogonal. If they are, is $\mathbb{R}^4 = U \oplus V$, that is, is $U^\perp = V$?

2.88. 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) $u_1 = (t, t, t)$, $\quad u_2 = (-4t, -4t, 4t)$, $\quad u_3 = (-2, -2, -2)$.

2.89. Construct an orthogonal basis of the subspace

$$\langle\, (1, 1, 1, 1),\ (1, 1, 1, -1),\ (-1, 1, 1, 1)\,\rangle$$

of the space $\mathbb{R}^4$.

2.90. In the space $\mathbb{R}^4$ find some orthogonal basis of the subspace of all linear combinations of the vectors $(1, 0, 1, 0)$, $(0, 1, 0, -7)$, $(4, -2, 4, 14)$ and the subspace generated by the vectors $(1, 2, 2, -1)$, $(1, 1, -5, 3)$, $(3, 2, 8, -7)$.

2.91. 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?

eigenspaces $V_{\pm 1}$ associated to the real eigenvalues and restrict our 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 $\dim V = 2n > 0$.

Let us now choose some eigenvalue λ and let u_λ be the eigenvector associated to the eigenvalue $\lambda = \alpha + i\beta$, $\beta \neq 0$. Analogously to the case of rotation in the plane given in the paragraph 2.45 by the matrix D we are interested in the real part of the sum of two one-dimensional subspaces $\langle u_\lambda\rangle \oplus \langle \bar{u}_\lambda\rangle$, where $\bar{u}_\lambda$ is the eigenvector associated to the eigenvalue $\bar{\lambda}$.

It is an intersection of the given sum of the complex subspaces with $\mathbb{R}^{2n}$, which is generated by the vectors $u_\lambda + \bar{u}_\lambda$ and $i(u_\lambda - \bar{u}_\lambda)$, that is, real vector subspace $P_\lambda \subset \mathbb{R}^{2n}$ generated by the basis given by the real and imaginary part of u_λ

$$x_\lambda = \operatorname{re} u_\lambda, \quad -y_\lambda = -\operatorname{im} u_\lambda.$$

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 sizes, the size of the eigenvalue λ must be equal to one. But that means that the restriction of our mapping on P_λ is rotation through 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 just have it expressed in basis x_λ, y_λ, that is, we must in the coordinates rotate through the angle with opposite sign.

The proof of the whole theorem is finished, because the by restriction of our mapping to the orthogonal complement and repeating the previous we obtain the whole decomposition after n steps. □

We return to the ideas in this proof once again in the third chapter, when we study complex extensions of the Euclidean vector spaces, see 3.26.

Remark. Specially in the dimension three at least one eigenvalue ±1 must be real, because three is an odd number. But then the associated eigenspace is an axis of the rotation of the three-dimensional space through the angle given by argument of the other eigenvalues. Try to think how to detect in which direction the space is rotated and also that the eigenvalue -1 means additional reflection through the plane perpendicular to the axis of the rotation.

We shall return to the discussion of the properties of matrices and linear mappings. Before we continue with the general theory, we show first in the next chapter a couple of application. We close this section with a general definition:

Spectrum of linear mapping

2.51. Definition. *Spectrum of linear mapping* $f : V \to V$ (spectrum of a matrix) is a sequence of roots of the characteristic polynomial f, along with multiplicities. *Algebraic multiplicity* of eigenvalue means its multiplicity as of the root of the characteristic polynomial, *geometric multiplicity* of the eigenvalue is the dimension of the associated subspace of eigenvectors.

Spectral diameter of of a linear mapping (of a matrix) is the greatest of the absolute values of the eigenvalues.

2.92. 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 some basis of its orthogonal complement.

2.93. Describe the orthogonal complement of the subspace V of the space $\mathbb{R}^4$, if V is generated by the vectors $(-1, 2, 0, 1)$, $(3, 1, -2, 4)$, $(-4, 1, 2, -4)$, $(2, 3, -2, 5)$.

2.94. 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 $x_1 - x_3 = 0$, $x_1 - x_2 + x_3 - x_4 + x_5 = 0$.

2.95. Let in the space $\mathbb{R}^4$ be given vectors

$$(1, -2, 2, 1), \quad (1, 3, 2, 1).$$

Extend these two vectors into an orthogonal basis of the whole $\mathbb{R}^4$. (You can do it in any way you like, for instance using the Gram-Schmidt orthogonalisation process.)

2.96. Find some orthonormal basis of the subspace $V \subset \mathbb{R}$, where $V = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \,|\, x_1 + 2x_2 + x_3 = 0\}$.

Solution. We see that the fourth coordinate does not appear in the restriction for the subspace, thus it seems reasonable to pick $(0, 0, 0, 1)$ as one of the vectors of the orthonormal basis and reduce the problem into the subspace $\mathbb{R}^3$. Let us try once again to avoid any computation – we see that 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 opposite value of the sum of the first and the third coordinate, we then normalise, that is, we choose the vector $(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, 0)$ and we are done. □

J. Eigenvalues and eigenvectors

2.97. Eigenvalues and eigenvectors can be used to illustrative description of linear mappings, notably in $\mathbb{R}^2$ and $\mathbb{R}^3$.

(1) Consider mapping with a matrix under the standard basis

$$f : \mathbb{R}^3 \to \mathbb{R}^3,\ A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

In this terminology, our results about orthogonal mappings can be formulated as follows: the spectrum of orthogonal mapping is always a subset of the unit circle in the complex plane. That means that in the real part of the spectrum there are only values ±1, whose algebraic and geometric multiplicities are the same. Complex values of the spectrum then correspond to the rotations in suitable two-dimensional subspaces which are mutually perpendicular.

We then obtain

$$|A - \lambda E| = \begin{vmatrix} -\lambda & 0 & 1 \\ 0 & 1-\lambda & 0 \\ 1 & 0 & -\lambda \end{vmatrix} = -\lambda^3 + \lambda^2 + \lambda - 1,$$

with roots $\lambda_{1,2} = 1, \lambda_3 = -1$. 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 thanks to 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, $\mathbb{R}^3 = V_1 \oplus V_2$, $\dim V_1 = 2$, $\dim V_2 = 1$. This decomposition is uniquely determined and says a lot about geometric properties of the mapping f. The eigenspace V_1 is furthermore a direct sum of one-dimensional eigenspaces, which can be chosen in more ways (thus such a decomposition has no further geometrical meaning).

(2) Consider 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

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

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.

2.98. An exercise with a change of basis. 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 under the basis:

$$\begin{aligned} e_1 &= [1, -1, 1] \\ e_2 &= [1, 2, 0] \\ e_3 &= [0, 1, 1] \end{aligned}$$

Solution. Characteristic polynomial of the matrix is

$$\begin{vmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 2 & 1-\lambda \end{vmatrix} = -\lambda^3 + 4\lambda^2 - 2\lambda = -\lambda(\lambda^2 - 4\lambda + 2).$$

Roots of this polynomial, eigenvalues, say when the matrix

$$\begin{pmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 2 & 1-\lambda \end{pmatrix}$$

will not have full rank, that is, the system of equations

$$\begin{pmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 2 & 1-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

will have more solutions than just $\mathbf{x} = (0, 0, 0)$. Thus eigenvalues are $0, 2 + \sqrt{2}, 2 - \sqrt{2}$. Let us compute 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 one-dimensional vector space of eigenvectors: $\langle(1, -1, 1)\rangle$.

- $2 + \sqrt{2}$: 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 $\langle(1, 1+\sqrt{2}, 1+\sqrt{2})\rangle$.

- $2 - \sqrt{2}$: 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 $\langle(1, 1-\sqrt{2}, 1-\sqrt{2})\rangle$.

The given matrix has eigenvalues 0, $2+\sqrt{2}$ and $2-\sqrt{2}$, with associated one-dimensional spaces of eigenvectors $\langle(1,-1,1)\rangle$, $\langle(1,1+\sqrt{2},1+\sqrt{2})\rangle$ and $\langle(1,1-\sqrt{2},1-\sqrt{2})\rangle$ 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 under 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 2.38)

$$B=TAT^{-1}=\begin{pmatrix}0&5&2\\0&-2&-1\\0&14&6\end{pmatrix}.$$

□

Let us do some more exercises for computing with eigenvalues and eigenvectors.

2.99. 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. Let us first construct 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 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 (that is, the so-called geometric multiplicity of the eigenvalue is also one, see 3.32).

Let us determine the eigenvector associated with the eigenvalue 2 (it is a solution of the homogeneous linear system with the matrix $A - 2E$):

$$\begin{aligned} -3x_1 + x_2 &= 0 \\ -1x_1 + x_2 &= 0 \\ 2x_1 - 2x_2 &= 0. \end{aligned}$$

The system has solution $x_1 = x_2 = 0, x_3 \in \mathbb{R}$ arbitrary, the eigenvector associated to the value 2 is then for instance the eigenvector (and any multiple of it).

Analogically we determine the remaining two eigenvectors – as solutions of the system $[A - (1 + \sqrt{3})E]\mathbf{x} = 0$ and 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 $\{\left((\frac{\sqrt{3}}{2} - 1)t, -\frac{t}{2}, \right), \ t \in \mathbb{R}\}$. That is the space of eigenvectors associated with the eigenvalue $1 + \sqrt{3}$ (except for the zero vector, which is a solution of the system, but we do not consider it an eigenvector; we shall not refer to this anymore in the future and we won't explicitly exclude the zero vector from the set of solutions).

Similarly we obtain that the space of eigenvector associated with the eigenvalue $1 - \sqrt{3}$ is $\langle(-1 - \frac{\sqrt{3}}{2}, -\frac{1}{2}, 1)\rangle$. □

2.100. 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. Characteristic polynomial of the matrix is $\lambda^3 - 6\lambda^2 + 12\lambda - 8$, which is $(\lambda - 2)^2$ with a root 2 which has multiplicity 3. The number 2 is thus an eigenvalue with algebraic multiplicity three. Its geometric multiplicity is either one, two or three. Let us determine the vectors associated to this eigenvalue as the solutions of the system

$$(A - 2E)\mathbf{x} = \begin{array}{rrcl} -x_1 & +x_2 & = & 0, \\ -x_1 & +x_2 & = & 0, \\ 2x_1 & -2x_2 & = & 0. \end{array}$$

Its solutions form the two-dimensional space $\langle(1, -1, 0), (0, 0, 1)\rangle$. The eigenvalue 2 has thus algebraic multiplicity 3 and geometric multiplicity 2.

□

Further basic exercises regarding eigenvalues and eigenvectors of matrices can be found at the page 130

2.101. 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

$$A = \begin{pmatrix} 32 & -67 & 47 \\ 7 & -14 & 13 \\ -7 & 15 & -6 \end{pmatrix}.$$

○

2.102. Without any computation write down the spectrum of the linear mapping $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by $(x_1, x_2, x_3) \mapsto (x_1 + x_3, x_2, x_1 + x_3)$.

○

2.103. Give the dimension of the eigenspaces of the eigenvalues λ_i of the matrix

$$\begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 5 & 2 & 3 & 0 \\ 0 & 4 & 0 & 3 \end{pmatrix}.$$

○

2.104. Pauli matrix In physics, state of the 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 it holds 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 eigenvalues of the matrices $\sigma_1, \sigma_2, \sigma_3$ are ±1.

Show that for matrices describing the state of the particle with spin 1

$$\frac{1}{\sqrt{2}}\begin{pmatrix}0 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 0\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}0 & -i & 0\\ i & 0 & -i\\ 0 & i & 0\end{pmatrix}, \begin{pmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & -1\end{pmatrix}$$

the same commuting relations hold as in the case of Pauli matrices.

Equivalently it can be shown that under the notation $I := \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$, $I := i\sigma_3$, $J := i\sigma_2$, $K := i\sigma_1$ forms the vectors space with basis $(1, I, J, K)$ an algebra of quaternions (algebra is a vector space with binary bilinear operation of multiplication, in this case the multiplication is given by the matrix multiplication). In order for the vector space to be an algebra of quaternions it is necessary and sufficient to show the following properties: $I^2 = J^2 = K^2 = -I$ and $IJ = -JI = K$, $JK = -KJ = I$ and $KI = -IK = J$.

2.105. 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 it is possible, give an example of such tuple of matrices D, P, and find out how many such tuples are there. ○

As we have already seen in , based on the eigenvalues and eigenvectors of the given 3×3 matrix, we can often geometrically interpret the mapping it gives in $\mathbb{R}^3$. Notably, we can do it in these situations:

If the matrix has 0 as eigenvalue and 1 as an eigenvalue with geometric multiplicity 2, 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, 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, it is a mirror symmetry through the plane of the eigenvectors associated with 1.

If the matrix has eigenvalue 1 with eigenvector perpendicular to plane of the eigenvectors associated with the eigenvalue -1, it is an axial symmetry (in space) through the axis given by the eigenvector associated with 1.

2.106. 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 $\langle(2,0,1),\ (1,1,0)\rangle$. 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 $x+4y-3z=0$. □

2.107. The theorem (2.50) gives us tools how to recognise matrix of a rotation in $\mathbb{R}^3$: 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 $u_\lambda + \overline{u_\lambda}$, $i[u_\lambda - \overline{u_\lambda}]$.

2.108. 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 the 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, $(1,1+i,-1-i)$; $\frac{3}{5}-\frac{4}{5}i$, $(1,1-i,-1+i)$. It is thus a matrix of a rotation (all the eigenvalues have absolute value 1 and one of the eigenvalues is 1), further we know that it is a rotation by the angle $\arccos(\frac{3}{5}) \doteq 0,295\pi$, which is the argument of the complex number $\frac{3}{5}+\frac{4}{5}i$. It remains to determine the direction of the rotation. First it is good to 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 2.50, 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 $((0,1,-1),(1,1,-1))$. 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 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 $(0, 1, 1)$. □

K. Additional exercises for the whole chapter

2.109. Solve the equation

$$\begin{array}{rcrcrcrcrcr} x_1 & + & x_2 & + & x_3 & + & x_4 & - & 2x_5 & = & 3, \\ & & 2x_2 & + & 2x_3 & + & 2x_4 & - & 4x_5 & = & 5, \\ -x_1 & - & x_2 & - & x_3 & + & x_4 & + & 2x_5 & = & 0, \\ -2x_1 & + & 3x_2 & + & 3x_3 & & & - & 6x_5 & = & 2. \end{array}$$

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. We thus consider the system

$$\begin{array}{rcrcrcrcrcr} x_1 & + & x_2 & + & t & + & x_4 & - & 2s & = & 3, \\ & & 2x_2 & + & 2t & + & 2x_4 & - & 4s & = & 5, \\ & & & & & & 2x_4 & & & = & 3. \end{array}$$

We see that $x_4 = 3/2$. The second equation gives

$$2x_2 + 2t + 3 - 4s = 5, \qquad \text{that is,} \qquad x_2 = 1 - t + 2s.$$

>From the first we have

$$x_1 + 1 - t + 2s + t + 3/2 - 2s = 3, \qquad \text{tj.} \quad x_1 = 1/2.$$

Altogether,

(2.1)

$$(x_1,\ x_2,\ x_3,\ x_4,\ x_5) = (1/2,\ 1 - t + 2s,\ t,\ 3/2,\ s), \qquad t, s\ \in \mathbb{R}.$$

In this exercise we also consider the extended matrix and we transform it using the row transformations into the row echelon form, where the first non-zero number in every row is 1 and where in a column in which this 1 is located the remaining numbers are 0. We also note that we omit the fourth equation, which is a combination of the first three. Gradually, 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

$$\left(\begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 2 & 2 & 2 & -4 & 5 \\ 0 & 0 & 0 & 2 & 0 & 3 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 1 & 1 & 1 & -2 & 5/2 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array}\right) \sim$$

$$\left(\begin{array}{ccccc|c} 1 & 1 & 1 & 0 & -2 & 3/2 \\ 0 & 1 & 1 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & 1 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array}\right).$$

If we choose again $x_3 = t$, $x_5 = s$ ($t, s \in \mathbb{R}$), we obtain the general solution ($\|2.1\|$) in the same form, directly. Consider the corresponding equations

$$\begin{array}{rcrcrcrcl} x_1 & & & & & & & = & 1/2, \\ & x_2 & + & t & & - & 2s & = & 1, \\ & & & & x_4 & & & = & 3/2. \end{array}$$

□

2.110. Find the solution of the system of linear equations given by the extended matrix

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 & 4 \\ 0 & 5 & -4 & 3 & 1 \\ 5 & 3 & 3 & -3 & 5 \end{array}\right).$$

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

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 & 4 \\ 0 & 5 & -4 & 3 & 1 \\ 5 & 3 & 3 & -3 & 5 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 5 & -4 & 3 & 1 \\ 0 & 6 & 1 & 14 & 0 \end{array}\right).$$

Copying the first two row s and adding a 5-multiple of the second row to the 3-multiple of the third and its 2-multiple to the fourth gives

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 5 & -4 & 3 & 1 \\ 0 & 6 & 1 & 14 & 0 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & -17 & -1 & 33 \\ 0 & 0 & -1 & 10 & 12 \end{array}\right).$$

We copy the first, second and fourth row, and add the fourth to the third, we obtain

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & -17 & -1 & 33 \\ 0 & 0 & -1 & 10 & 12 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & -18 & 9 & 45 \\ 0 & 0 & -1 & 10 & 12 \end{array}\right).$$

Then we have (the remaining row transformations are „usual“)

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & -18 & 9 & 45 \\ 0 & 0 & -1 & 10 & 12 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & 2 & -1 & -5 \\ 0 & 0 & 1 & -10 & -12 \end{array}\right) \sim$$

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & 1 & -10 & -12 \\ 0 & 0 & 2 & -1 & -5 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & 1 & -10 & -12 \\ 0 & 0 & 0 & 19 & 19 \end{array}\right).$$

We see that the system has exactly 1 solution. We determine it by backwards elimination

$$\left(\begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 0 & -3 & -1 & -2 & 6 \\ 0 & 0 & 1 & -10 & -12 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cccc|c} 3 & 3 & 2 & 0 & 2 \\ 0 & -3 & -1 & 0 & 8 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{cccc|c} 3 & 3 & 0 & 0 & 6 \\ 0 & -3 & 0 & 0 & 6 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 4 \\ 0 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right).$$

The result is then

$$x_1 = 4, \quad x_2 = -2, \quad x_3 = -2, \quad x_4 = 1.$$

□

2.111. Give 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 at x, in the second there are coefficients at y, and so on, while 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 can directly obtain all solutions

$$\begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = t \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -1 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \quad t, s \in \mathbb{R},$$

because we have the matrix in the row echelon form, while the first non-zero number in every row is 1 and in a column where there is such 1 there are zeroes at all other positions. The solution given as a linear combination of two vectors is determined exactly by the columns without first non-zero number on some row, that is, by the second and the fifth column, when we choose 1 as the second coordinate for the second column and as the fifth coordinate for the fifth column and when we take the numbers in the corresponding column with the opposite sign and put them at the position given by the column where there is a first 1 in its row. Let us add that the result can be immediately rewritten in the form

$$(x,\ y,\ z,\ u,\ v) = (-t - s,\ t,\ -s,\ 0,\ s)\ , \quad t, s \in \mathbb{R}.$$

□

2.112. Decompose the following permutations into transposition:

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) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 6 & 1 & 10 & 2 & 5 & 9 & 8 & 3 & 7 \end{pmatrix}$.

2.113. 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) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 7 & 1 & 10 & 2 & 5 & 4 & 9 & 3 & 6 \end{pmatrix}$.

2.114. Determine the eigenvalues of the matrix

$$\begin{pmatrix} -13 & 5 & 4 & 2 \\ 0 & -1 & 0 & 0 \\ -30 & 12 & 9 & 5 \\ -12 & 6 & 4 & 1 \end{pmatrix}.$$

○

2.115. Having been told that the numbers 1, −1 are the eigenvalues of the matrix

$$A = \begin{pmatrix} -11 & 5 & 4 & 1 \\ -3 & 0 & 1 & 0 \\ -21 & 11 & 8 & 2 \\ -9 & 5 & 3 & 1 \end{pmatrix},$$

give all solutions of the characteristic equation $|\, A - \lambda\, E \,| = 0$. Hint: if you denote all the roots of the polynomial $|\, A - \lambda\, E \,|$ as $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, it is

$$|\, A \,| = \lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \lambda_4, \quad \operatorname{tr} A = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4.$$

○

2.116. Give an example of a four-dimensional matrix with eigenvalues $\lambda_1 = 6$ and $\lambda_2 = 7$ such that their 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;

○

2.117. Find the eigenvalues and eigenvectors of the matrix:

$$\begin{pmatrix} -1 & -\frac{5}{6} & \frac{5}{3} \\ 0 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{1}{6} & -\frac{4}{3} \end{pmatrix}.$$

2.118. 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

2.7.

$$A^5 = \begin{pmatrix} 122 & -121 & 121 \\ -121 & 122 & -121 \\ 0 & 0 & 1 \end{pmatrix}, \quad A^{-3} = \tfrac{1}{27}\begin{pmatrix} 14 & 13 & -13 \\ 13 & 14 & 13 \\ 0 & 0 & 27 \end{pmatrix}.$$

2.12. There is only one such matrix X, and it is

$$\begin{pmatrix} 18 & -32 \\ 5 & -8 \end{pmatrix}.$$

2.14. $A^{-1} = \begin{pmatrix} 1 & 10 & -4 \\ 1 & 12 & -5 \\ 0 & 5 & -2 \end{pmatrix}.$

2.15. $\begin{pmatrix} 2 & -3 & 0 & 0 & 0 \\ -5 & 8 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -5 & 2 \\ 0 & 0 & 0 & 3 & -1 \end{pmatrix}.$

2.16. $C^{-1} = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$

2.17. In the first case we have

$$A^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 3 & -i \\ i & 1 \end{pmatrix};$$

in the second

$$A^{-1} = \begin{pmatrix} 14 & 8 & 5 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

2.18. We have

$$A^{-1} = \frac{1}{n-1}\begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}.$$

2.21. -3,17,-1

2.24. By subtracting the first row from all other rows and then expanding the first column we obtain

$$\begin{aligned} V_n(x_1, x_2, \dots, x_n) &= \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix} \\ &= \begin{vmatrix} x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix}. \end{aligned}$$

If we take out $x_{i+1} - x_1$ from the i-th row for $i \in \{1, 2, \dots, n-1\}$, we obtain

$$\begin{aligned} &V_n(x_1, x_2, \dots, x_n) \\ &= (x_2 - x_1) \cdots (x_n - x_1) \begin{vmatrix} 1 & x_2 + x_1 & \dots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \dots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix}. \end{aligned}$$

By subtracting from every column (starting with the last and ending with the second) x_1-multiple of the previous column, we obtain

$$\begin{vmatrix} 1 & x_2 + x_1 & \dots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \dots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix} = \begin{vmatrix} 1 & x_2 & \dots & x_2^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-2} \end{vmatrix}.$$

Therefore

$$V_n(x_1, x_2, \dots, x_n) = (x_2 - x_1) \cdots (x_n - x_1)\ V_{n-1}(x_2, \dots, x_n).$$

Because it is clear that

$$V_2(x_{n-1}, x_n) = x_n - x_{n-1},$$

it holds (by induction) that

$$V_n(x_1, x_2, \dots, x_n) = \prod_{1 \le i < j \le n} (x_j - x_i).$$

Note that the determinant is non-zero whenever the numbers $x_1, \dots, x_n$ are mutually distinct.

2.27.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}^{-1} \quad \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

We can then easily obtain

$$x_1 = \frac{13}{4}, \quad x_2 = -\frac{3}{4}, \quad x_3 = -\frac{3}{4}, \quad x_4 = \frac{1}{4}.$$

2.33. The solutions are exactly all scalar multiples of a vector

$$\left(1 + \sqrt{3},\ -\sqrt{3},\ 0,\ 1,\ 0\right).$$

2.34. $x_1 = 1 + t, \quad x_2 = \frac{3}{2}, \quad x_3 = t, \quad x_4 = -\frac{1}{2}, \qquad t \in \mathbb{R}.$

2.35. The system has no solution.

2.36. The system has a solution, because

$$3 \cdot \begin{pmatrix} 3 \\ 2 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3 \\ -3 \\ -2 \end{pmatrix} - 5 \cdot \begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \\ 4 \\ 6 \end{pmatrix}.$$

2.37. System of linear equations

$$\begin{array}{rcrcrcl} 3x_1 & & & + & 2x_3 & = & 1, \\ x_1 & & & + & x_3 & = & 2, \\ 7x_1 & & & + & 4x_3 & = & 3, \\ 5x_1 & & & + & 3x_3 & = & 4, \\ & & x_2 & & & = & 5 \end{array}$$

has no solution, while the system

$$\begin{array}{rcrcrcl} 3x_1 & & & + & 2x_3 & = & 1, \\ x_1 & & & + & x_3 & = & 1, \\ 7x_1 & & & + & 4x_3 & = & 1, \\ 5x_1 & & & + & 3x_3 & = & 1, \\ & & x_2 & & & = & 1 \end{array}$$

has a unique solution $x_1 = -1$, $x_2 = 1$, $x_3 = 2$.

2.38. The set of all solutions is

$$\{(-10t,\ (a+4)t,\ (3a-8)t)\ ;\ t \in \mathbb{R}\}.$$

2.39. For $a = 0$ the system has no solution, for $a \neq 0$ it has infinitely many solutions.

2.40. The correct answers are „yes“, „no“, „no“ and „yes“ respectively.

2.41. i) For $b \neq -7$ is $x = z = (2+a)/(b+7)$, $y = (3a - b - 1)/(b+7)$ (1b). ii) For $b = -7$ (1b) and $a \neq -2$ (1b) has no solution (1b), for $a = -2$ the solution is $x = z$, $3z - 1$ (2b).

2.43. 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 $\alpha, \beta, \gamma, \delta \in \mathbb{R}$.

2.44. The matrices are

$$\text{(a)} \begin{pmatrix} 1 & 1 & -2 & -4 \\ 0 & 1 & 0 & -1 \\ -1 & -1 & 3 & 6 \\ 2 & 1 & -6 & -10 \end{pmatrix}, \qquad \text{(b)} \begin{pmatrix} 6 & -2i \\ -3+2i & 1+i \end{pmatrix}.$$

2.47. 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.

2.52. There is a unique solution

$$p = 2, \qquad q = -2, \qquad r = 3.$$

2.55. $(2 + \frac{1}{\sqrt{3}}, 2 - \frac{1}{\sqrt{3}})$.

2.56. The vectors are dependent whenever at least one of the conditions

$$a = b = 1, \qquad a = c = 1, \qquad b = c = 1$$

is satisfied.

2.57. Vectors are linearly independent.

2.58. It suffices to add for instance the polynomial x.

2.70.

$$\begin{pmatrix} 1/4 & -\sqrt{6}/4 & 3/4 \\ \sqrt{6}/4 & -1/2 & -\sqrt{6}/4 \\ 3/4 & \sqrt{6}/4 & 1/4 \end{pmatrix}$$

2.74.

$$\begin{pmatrix} 5/6 & -1/6 & 1/3 \\ -1/6 & 5/6 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

2.75.

$$\begin{pmatrix} 5/9 & 2/9 & -4/9 \\ 2/9 & 8/9 & 2/9 \\ -4/9 & 2/9 & 5/9 \end{pmatrix}$$

2.80. $\cos = \frac{\sqrt{2}}{\sqrt{3}}$.

2.81.

$$q_1 : (2 - \frac{\sqrt{3}}{2}, 2\sqrt{3} + \frac{1}{2})t, \qquad q_2 : (2 + \frac{\sqrt{3}}{2}, -2\sqrt{3} + \frac{1}{2})t.$$

2.84. For instance 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

$$\left\| \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix} \right\| = \left\langle \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 1 & -2 & -3 \end{pmatrix} \right\rangle = \sqrt{1^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 1^2 + (-2)^2 + (-3)^2} = \sqrt{23}.$$

2.87. The vector that gives 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

$$(-1, 0, -1, 2) = (-1, 0, 1, 0) - 2\,(0, 0, 1, -1)\,.$$

2.88. 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 $t = 0$.

2.89. Using the Gram-Schmidt orthogonalisation process we can obtain the result

$$((1, 1, 1, 1),\ (1, 1, 1, -3),\ (-2, 1, 1, 0))\,.$$

2.90. Preserving the order of the subspaces from the problem statement we have for instance the orthogonal bases

$$((1, 0, 1, 0), (0, 1, 0, -7))$$

and

$$((1, 2, 2, -1), (2, 3, -3, 2), (2, -1, -1, -2)).$$

2.91. The result is $a = 9/2$, $b = -5$, because it must hold

$$1 + b + 4 + 0 + 0 = 0, \quad 1 - b + 0 + 3 - 2a = 0.$$

2.92. The basis must contain a single vector. It is some non-zero scalar multiple of the vector

$$(3, -7, 1, -5, 9).$$

2.93. Orthogonal complement $V^\perp$ is a set of all scalar multiples of the vector $(4, 2, 7, 0)$.

2.94.

(a) $W^\perp = \langle\, (1, 0, -1, 1, 0),\ (1, 3, 2, 1, -3)\,\rangle\,;$
(b) $W^\perp = \langle\, (1, 0, -1, 0, 0),\ (1, -1, 1, -1, 1)\,\rangle.$

2.95. There is infinitely many possible extensions, of course. A very simple one is for instance

$$(1, -2, 2, 1), \quad (1, 3, 2, 1), \quad (1, 0, 0, -1), \quad (1, 0, -1, 1).$$

2.101. Je $|\,A - \lambda\,E\,| = -\lambda^3 + 12\lambda^2 - 47\lambda + 60,$. $\lambda_1 = 3$, $\lambda_2 = 4$, $\lambda_3 = 5$.

2.102. The result is the sequence 0, 1, 2.

2.103. Dimension is 1 for $\lambda_1 = 4$ and 2 for $\lambda_2 = 3$.

2.105. Matrix B has two distinct eigenvalues, and thus such 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 tuples D, P.

2.112. i) $(1, 7)(2, 6)(5, 3)$, ii) $(1, 6)(6, 8)(8, 7)(7, 3)(2, 4)$, iii) $(1, 4)(4, 10)(10, 7)(7, 9)(9, 3)(2, 6)(6, 5)$

2.113. i) 17 inversions, odd, ii) 12 inversions, even iii) 25 inversions, odd

2.114. The matrix has only one eigenvalue – -1.

2.115. The root -1 of the polynomial $|\,A - \lambda\,E\,|$ has multiplicity three.

2.116. For instance,

$$\text{(a)}\ \begin{pmatrix}6&0&0&0\\0&7&0&0\\0&0&7&0\\0&0&0&7\end{pmatrix};\quad \text{(b)}\ \begin{pmatrix}6&0&0&0\\0&7&1&0\\0&0&7&0\\0&0&0&7\end{pmatrix};$$

$$\text{(c)}\ \begin{pmatrix}6&0&0&0\\0&7&1&0\\0&0&7&1\\0&0&0&7\end{pmatrix}.$$

2.117. Triple eigenvalue -1, corresponding eigenspace is $\langle(1,0,0),(0,2,1)\rangle$.

2.118. Characteristic polynomial is $-(\lambda-2)^2(\lambda-9)$, that is, eigenvalues are 2 and 9 with associated eigenvectors

$$(1,2,0)\,,(-3,0,1)\quad \text{a}\quad (1,1,1)$$

CHAPTER 3

Linear models and matrix calculus

where are the matrices useful?

– basically almost everywhere…

A. Processes with linear restrictions

Let us show an example of a very simple linear optimisation problem:

3.1. A company manufactures bolts and nuts. Nuts and bolts are moulded – moulding a box of bolts takes one minute, 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 matrices. Due to technical reasons it is not possible to manufacture more than 110 boxes of bolts. The profit from one box of bolts is 40 Kč and the profit from one box of 60 Kč. The company has no trouble with selling. How many boxes of nuts and bolts should be manufactured in order to have maximal profit?

Solution. Let us 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	40 Kč/box	60 Kč/box	

We have already developed a pretty useful package of tools and it is time to show some applications of the matrix calculus. On some relatively easy problems we see that the theory allows us both qualitative and quantitative analyses and sometimes it leads quite easily to some surprising results.

Although it might seem that the assumption of linearity of relations between the quantities is too restrictive, it is quite often note so – in real problems there linear relations tend to either directly appear or the final process is a result of an iteration of many linear steps. And even if it is not the case, we can using this approach at least approximate the real processes.

We like to view the matrices (and linear mappings) as objects with which we would like to work with as if they were scalars. In order to do that, a pretty hard work to be done in the fourth chapter is required. We show a quick and useful application then on the so-called matrix decompositions, which are needed for numerical mastery of matrix calculus in a most robust way.

1. Linear processes

3.1. Solution of system of linear equations. Simple linear processes are given by linear mappings $\varphi : V \to W$ 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 was realised.

If we want to reach a given result $b \in W$ of such process, we solve the problem

$$\varphi(x) = b$$

for some unknown vector x and a known vector b.

In fixed coordinates we then have a matrix A of a mapping φ and coordinate expression of the vector b. As we have already noted in the introduction to the second chapter, the set of all solutions of the so-called *homogeneous system*

$$A \cdot x = 0$$

is a vector space.

If the dimension of V is finite, say n, and the dimension of the image of the mapping φ is k, then by solving of this system using the row echelon transformation (see 2.7) we find out that the dimension of the space of all solutions is exactly $n - k$. Indeed, as the columns of the matrix of the mapping are exactly the images of the vectors of the basis, in the matrix of the system there are exactly k linearly independent columns and thus also the same number of linearly independent rows. Therefore even after doing the transformation into the row echelon form exactly $n - k$ zero rows remain.

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 the moulding time and from the restriction on the box preparation we obtain the following restrictive conditions:

$$\begin{aligned} x_1 + 2x_2 &\leq 120 \\ x_1 + 4x_2 &\leq 180 \\ x_1 &\geq x_2 + 90 \\ x_1 &\geq 110 \end{aligned}$$

The objective function (the function that gives the profit for given number of manufactured nuts and bolts) is $40x_1 + 60x_2$. The previous system of inequalities gives $\mathbb{R}^2$ a certain area and the optimisation of the profit means to find in this area the point (points) in which the objective function has the maximum value, that is, to find the highest k such that the line $40x_1 + 60x_2 = k$ has a non-empty intersection with the given area. Graphically, we can find the solution for example by placing the line p into the plane such that it satisfies the equation $40x_1 + 60x_2 = 0$ 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 the borderline of the line (and the line must be parallel to p). Thus we obtain (see the figure) the point $x_1 = 110$ and $x_2 = 5$. Maximum possible income is thus $40 \cdot 100 + 60 \cdot 5 = 4700$ Kč. □

3.2. **Minimisation of costs for feeding** Stable in Nišovice u Volyně buys fodder for winter: hay and oat. The nutritional values of the fodder and required daily portions for one foal are given in the table:

g/kg	Hay	Oat	Requirements
Dry basis	841	860	At least 6300 g
Digestible nitrogen stuff	53	123	At most 1150 g
Starch	0,348	0,868	At most 5,35 g
Calcium	6	1,6	At least 30 g
Phosphate	2,8	3,5	At most 44 g
Natrium	0,2	1,4	Approximately 7 g
Cost	1,80	1,60	

Every foal must obtain in daily meal at least 2 kg of oat. Average cost (counting the payment for the transportation) cost $1,80$ Kč per 1 kg of hay and $1,60$ Kč per 1 kg of oat. Compose daily diet for one foal which has minimum costs.

3.3. **Optimal distribution of material.** On inner wooden panelling of a cottage there are following requirements

- at most 120 planks of length 35 cm,
- from 180 to 330 planks of length 120 cm,
- at least 30 planks of length 95 cm.

When solving the system of equation we are thus left with exactly $n - k$ free parameters and by setting one of them to have the value one and making the other to be zero we obtain exactly $n - k$ linearly independent solutions. All solutions are then given by all the linear combinations of these $n - k$ solutions. Every such $(n - k)$-tuple of solutions is called *fundamental system of solutions* of the given homogeneous system of equations. We have proved:

Theorem. *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 of $\mathbb{K}^n$ of dimension $n - k$. Every basis of such subspace forms a fundamental system of solutions of the given homogeneous system.

3.2. Non-homogeneous systems of equations. Consider now the general system of equations

$$A \cdot x = b.$$

Let us now realise once again that the columns of the matrix A are actually images of the vectors of the standard basis in $\mathbb{K}^n$ under the linear mapping φ corresponding to the matrix A. If there is to be a solution, b must be in the image under φ and thus it must be a linear combination of the columns in A.

If we extend the matrix A by the column b, the number of linearly independent columns and thus also rows might increase (but does not have to). If this number increases, then b is not in the image and the system of equations does not have a solution. If on the other hand the number of linearly independent rows does not change after adding the column b to the matrix A, it means that b must be a linear combination of the columns of A. Coefficients of such combinations are then exactly the solutions of our system.

Consider now two fixed solutions x and y of our system and some solution z of the homogeneous system with the same matrix. Then clearly

$$\begin{aligned} A \cdot (x - y) &= b - b = 0 \\ A \cdot (x + z) &= 0 + b = b. \end{aligned}$$

Thus we can summarise:

3.3. Theorem. *The solution of 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 case the space of all solution 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.*

In literature, this theorem is often called *Frobenius theorem* and its usual formulation is "system has a solution if and only if the rank of its matrix equals the rank of its extended matrix".

3.4. Optimisation linear models. In the parallel column we have started this chapter with painting problems. We shall continue with this. Imagine that our very specialised painter in the black-white world is willing to paint facades of either small family houses or of big public building, and that he uses only black and white colours. He can arbitrarily choose, in what range will he do x units of area of the first type or y units of the second type. Let us assume that his maximal workload in a given interval is L units of area, his clean income (that is, after subtracting the costs) is c_1 per a unit of area for small houses and

But only planks of length 4 meters can be bought. The total waste cannot be greater that 360 cm. Determine the minimal number of planks that can be bought (and how to cut them) in order to meet the requirements.

c_2 per unit of area for big 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, while for big buildings the corresponding values are w_2 and b_2.

If we sum it all into (in)equalities, we obtain the conditions

$$x_1 + x_2 \leq L \tag{3.1}$$

$$w_1x_1 + w_2x_2 \leq W \tag{3.2}$$

$$b_1x_1 + b_2x_2 \leq B. \tag{3.3}$$

The total clean income of the painter,

$$h(x_1, x_2) = c_1x_1 + c_2x_2,$$

should be maximised.

Each of the given inequalities clearly gives in the plane of the variables (x_1, x_2) a half-plane, 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 the restrictions for the values (x_1, x_2) – either the restrictions are unsatisfiable, or they allow points inside of a polygon with at most five vertices, see the picture.

In general we speak of *linear programming problem* whenever we seek either maximum or minimum of a linear form h over $\mathbb{R}^n$ over a set bounded by a system of linear inequalities which we call *linear restrictions*. Vector on the right side is then called *vector of restrictions*, the linear form h is also called *objective function*.

Formulation with inequalities $\leq$ at restrictive conditions, non-negative variables and maximalisation of the objective functions is called *standard maximalisation problem*. On the other hand, *standard minimisation problem* is defined by seeking minimum of the objective function while the restrictive inequalities are $\geq$ and the variables are non-negative.

It is easy to see that every general linear programming problem can be transformed into a standard one of either type. Aside of sign changes we can work with decomposition of the variables that have no sign restriction into a difference of two non-negative ones. Without loss of generality we shall further work only with the standard maximisation problem.

How to solve such problem. We seek maximum of a linear form h over subsets M of a vector space which are given by linear inequalities, that is, in the plane by the intersection of half planes, in general we shall speak in the next chapter about half-spaces. 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 – that is, in the direction it either grows all the time or decreases. 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 parametrisation yields

$$t \mapsto h(u + t\,v) = h(u) + t\,h(v).$$

This expression is indeed with increasing parameter t always either increasing or decreasing, or constant (depending on whether $h(v)$ is positive, negative or zero).

Thus we surely must expect that problems similar to the one with the painter are either unsatisfiable (if the given set with restrictions is empty), or the profit is unbounded (if the restrictions give an unbounded of the space and the form h is in some of the unbounded directions non-zero) or they attain a maximal solution in at least one of the "vertices" of the set M (while usually that would be the case for only a single vector, it can also be the case that the maximum is attained on a part of the boundary of the area M.

3.5. Formulation using linear equations. Finding an optimum is not always as simple as in the previous case. The problem can contain many variables and restrictions and even deciding whether the set M of satisfiable points is non-empty can be a problem.

We don't have space here for the complete theory, but we mention at least two directions of ideas, which show that actually the solution can be always found in a way similar to the previous paragraph.

Let us begin by comparison with systems of linear equations – because we understand those well. Let us write the equations (3.1)-(3.3) in general form:

$$A \cdot x \leq b,$$

where x is now an n-dimensional vector, b is m-dimensional vector and A is the corresponding matrix. By an inequality between vectors we mean individual inequalities between coordinates. We want to maximise the product $c \cdot x$ for a given row vector of coefficients of the linear form h. If we add a new auxiliary variable 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

$$\begin{pmatrix} 1 & -c & 0 \\ 0 & A & E_m \end{pmatrix} \cdot \begin{pmatrix} z \\ x \\ x_s \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}$$

where the matrix is composed of the blocks with $1{+}n{+}m$ columns and $1+m$ rows, with corresponding individual components of the vectors. Additionally we require for all coordinates X and x_s non-negativity.

If the given system of equations has a solution, in this set of solutions we seek values for the variables z, x and x_s, such that all x are non-negative and z maximised. In the paragraph 4.11 on page 215 we will discuss this situation from the viewpoint of affine geometry.

Specifically, in our problem of black and white painter the system of linear equations looks like this:

$$\begin{pmatrix} 1 & -c_1 & -c_2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & w_1 & w_2 & 0 & 1 & 0 \\ 0 & b_1 & b_2 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} z \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ L \\ W \\ B \end{pmatrix}$$

3.6. Duality of linear programming. Consider the real matrix A with m rows and n columns, vector of restrictions b and row vector c giving the objective function. From these data we can compose two problems of linear programming for $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

Maximisation problem: Maximise $c \cdot x$ under the conditions $A \cdot x \leq b$ and $x \geq 0$.

Minimisation problem: Minimise $y^T \cdot b$ under the condition $y^T \cdot A \geq c^T$ and $y \geq 0$.

We say that these two problems are mutually dual. For deriving other properties of linear programming we first introduce some terminology.

We say that the problem is *solvable* if there is some *admissible vector* x which meets all restrictions. Solvable maximisation (minimisation) is *bounded*, if the objective function is bounded from above (bellow) over the set of admissible vectors.

Lemma. *If $x \in \mathbb{R}^n$ is an admissible vector for the standard maximisation problem and $y \in \mathbb{R}^m$ is admissible vector for the dual minimisation problem, then for the objective functions we have*

$$c \cdot x \leq y^T \cdot b$$

Proof. It is actually a simple observation: $x \geq 0$ and $c^T \leq y^T \cdot A$, but also $y \geq 0$ and $A \cdot x \leq b$, thus it must also hold that

$$c \cdot x \leq y^T \cdot A \cdot x \leq y^T \cdot b,$$

which is what we wanted to prove. □

>From here we immediately see 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 it holds that $c \cdot x = y^T \cdot b$, then both are optimal solution for the corresponding problem.*

3.7. Theorem (About duality)**.** *If a standard problem of linear programming is solvable and bounded, then its dual is also bounded and solvable, there exist an optimal solution for each of the problems and the optimal values of the corresponding objective functions are equal.*

Proof. One direction was already proved in the previous corollary. It remains to prove the existence of an optimal solution. That is easy to prove by constructing an algorithm, which we won't do in a great detail now. We will return to the missing part of the proof in the part about affine geometry at the page 215. □

Let us note yet another corollary of the just formulated duality theorem:

Corollary (Equilibrium theorem)**.** *Consider two admissible vectors x and y for the standard maximisation problem and its dual problem from the definition 3.6. Then both these 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. Consider that both relations regarding the zeroes among x_i and y_i hold. Then we can in the following computation calculate with equation, because the summands with strict inequality have zero coefficients anyway:

$$\sum_{i=1}^m y_i b_i = \sum_{i=1}^m y_i \sum_{j=1}^n a_{ij}x_j = \sum_{i=1}^m \sum_{j=1}^n y_i a_{ij} x_j$$

and from the same reason also

$$\sum_{i=1}^m \sum_{j=1}^n y_i a_{ij} x_j = \sum_{j=1}^n c_j x_j.$$

This shows one implication, thanks to the duality theorem.

Consider now that both x and y are optimal vectors. We thus know that

$$\sum_{i=1}^{m} y_i b_i \geq \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j \geq \sum_{j=1}^{n} c_j x_j,$$

but simultaneously the left- and right-hand sides are equal. Thus there is equality everywhere. If we rewrite the first equality as

$$\sum_{i=1}^{m} y_i \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) = 0$$

we see that it can be satisfied only if the relation from the statement holds, because it is a sum of non-negative numbers and it 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 zeroes among the additional variables and satisfying the restrictions.

3.8. Notes about linear models in economy. Our very schematic problem of black-white painter from the paragraph 3.4 can be used to illustrate one of the typical economical models, the so-called *model of production planing*. The model tries to capture the problem completely, that is, to capture both external and internal relations. Left-hand sides of the equations (3.1), (3.2), (3.3) and of the objective function $h(x_1, x_2)$ are expressing various production relations. Depending on the character of the problem, we have on the right-hand sides either exact values (and then we solve equations) or capacity restrictions and goal optimisation (then we obtain linear programming problems).

We can thus in general solve the problem of source allocation with supplier restrictions and either minimise costs or maximise income. We can also interpret duality from this point of view. If our painter would like to set up costs of his work y_L, of white colour y_W and of black colour y_B, then he minimises the objective function

$$L \cdot y_L + W y_W + B y_B$$

with restrictions

$$\begin{aligned} y_L + w_1 y_W + b_1 y_B &\geq c_1 \\ y_L + w_2 y_W + b_2 y_B &\geq c_2. \end{aligned}$$

But that is exactly the dual problem to the original one and the theorem 3.7 says that optimal state is such when the objective functions have the same value.

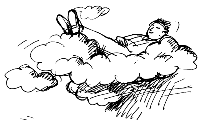
Among economical models we can find many modifications. One of them are problems of *financial planing*, which are connected to the optimisation of portfolio. We are setting up volume of investment into individual investment possibilities with the goal to meet the given restrictions for risk factors while maximising the profit, or dually minimise the risk under 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 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 show a more general theory for linear equations with constant coefficients, which gives not only very practical tools but also nice illustration for concepts of vector spaces and linear mappings.

Homogeneous linear difference equation of order k

3.9. Definition. *Homogeneous linear difference equation of order* k is given by the expression

$$a_0x_n + a_1x_{n-1} + \cdots + a_kx_{n-k} = 0, \quad a_0 \neq 0 \quad a_k \neq 0,$$

where the coefficients a_i are scalars, which can possibly also depend on n.

We also say that such equality gives *homogeneous linear recurrence* of order k and we usually denote the sequence in question as a function

$$x_n = f(n) = -\frac{a_1}{a_0}f(n-1) - \cdots - \frac{a_k}{a_0}f(n-k).$$

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 fixed n.

By giving any k values x_i in sequence are determined all the other values uniquely. Indeed, we work over a field of scalars, thus the values a_0 and a_k are invertible and thus using the definition any x_n can be computed uniquely, 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 coordinatewise. Directly from the definition is immediate that a sum of two solutions of a homogeneous linear 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 form a subspace.

Initial condition on the values of the solutions is given as a k-dimensional vector in $\mathbb{K}^k$. Sum of initial conditions determines the sum of the corresponding solutions, similarly for scalar multiples. Note also that plugging zeroes and ones into initial k values immediately yields k linearly independent solutions of the equation. Thus although the vectors are infinite sequences, the set of all solutions has finite dimension, we know that its dimension equals to the order of the equation k, and we can easily obtain a basis of all those solutions. Again we speak of *fundamental system of solutions* and all other solutions are its linear combinations.

As we have already checked, if we choose k indices i, $i + 1, \dots, i + k - 1$ in sequence, the homogeneous linear difference equation gives a linear mapping $\mathbb{K}^k \to \mathbb{K}^\infty$ of k-dimensional vectors of initial values into infinitely-dimensional sequences of the same scalars. Independence of such solutions is equivalent to the independence of the initial values – which can be easily told from determinant. If we have a k-tuple of solutions $(x_n^{[1]}, \dots, x_n^{[k]})$, it is independent if and only if the following determinant, the so-called *Casortian*, is non-zero for one n (which then implies it is non-zero for all n)

$$C(x_n^{[1]}, \dots, x_n^{[k]}) = \begin{vmatrix} x_n^{[1]} & \dots & x_n^{[k]} \\ x_{n+1}^{[1]} & \dots & x_{n+1}^{[k]} \\ \vdots & \ddots & \vdots \\ x_{n+k-1}^{[1]} & \dots & x_{n+k-1}^{[k]} \end{vmatrix} \neq 0$$

3.10. Solution of homogeneous recurrences with constant coefficients. It is hard to find a universal mechanism for finding a solution of general homogeneous linear difference equations, that is, directly computable expression for the general solution x_n.

In practical models there are very often equations, where the coefficients are constant. In this case it is possible to guess suitable form for the solution and indeed we will be able to find k linearly independent solutions. This is a complete solution of the problem, as all other solutions will be linear combinations.

For simplicity let us start with equations of second order. Such are very often encountered in practical problems, where there are relations based on two previous values. Linear difference equation (recurrence) of second order with constant coefficients is for us thus a form

$$f(n+2) = a \cdot f(n+1) + b \cdot f(n) + c, \tag{3.4}$$

where a, b, c are known scalar coefficients.

For instance in population models we can 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 $b \cdot f(n)$ with positive $b > 1$), while immature individuals tire and destroy part of the mature population (that is, the coefficient a is negative). Furthermore, it might be that somebody destroys (uses, eats) a fixed amount c every season.

Special such case with $c = 0$ is for instance the Fibonacci sequence of numbers $y_0, y_1, \dots$, where $y_{n+2} = y_{n+1} + y_n$.

If when solving a mathematical problem we don't have any new idea, we can always try to what success leads some known solution of a similar problem. Let us try to plug into the equation (3.4) with coefficient $c = 0$ similar solution as with the linear equations, that is, $f(n) = \lambda^n$ for some scalar λ. By plugging in 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

$$\lambda_1 = \frac{1}{2}(a + \sqrt{a^2 + 4b}), \quad \lambda_2 = \frac{1}{2}(a - \sqrt{a^2 + 4b}).$$

We have thus determined when actually such solutions indeed work, we just have to suitably choose the scalar λ. But this is not enough for us, since we need to find a solution for any two initial values $f(0)$ and $f(1)$, and we have only found two specific sequences satisfying given equation (or possibly only one sequence – if $\lambda_2 = \lambda_1$).

As we have already derived for even very general linear recurrences, sum of two solutions $f_1(n)$ and $f_2(n)$ of our equation $f(n+2) - a \cdot f(n+1) - b \cdot f(n) = 0$ is clearly again a solution

B. Recurrent equations

Distinct linear dependences can be a good tool for describing various models of growth. Let us begin with a very popular population model that uses linear difference equation of second order:

3.4. Fibonacci sequence. In the beginning of the Spring a stork brought on a meadow two newborn rabbits, male and female. The female is, after being two months old, 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 will be there after nine months (if none of them dies and none "moves 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*

$$p_{n+2} = p_{n+1} + p_n, \quad n = 1, \dots, \tag{3.1}$$

which along with initial conditions $p_1 = 1$ and $p_2 = 1$ uniquely determines the numbers of pairs of rabbits at the meadow in individual months. Linearity of the formula means that all members of the sequence (p_n) appear in the first power, the meaning of the word recurrence is hopefully clear and the homogeneity means that in the formula the absolute term is missing (see further for non-homogeneous formula). 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 plugging into the recurrent relation:

$$\begin{aligned} r^{n+2} &= r^{n+1} + r^n \quad \text{and after dividing by } r^n \text{ we obtain} \\ r^2 &= r + 1, \end{aligned}$$

which is the so-called *characteristic equation* of the giver recurrent formula. Our equation thus 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 of the same equation and the same holds for the scalar multiples of the solution. Our two specific solutions thus allow even more general solutions

$$f(n) = C_1\lambda_1^n + C_2\lambda_2^n$$

for arbitrary scalars C_1 and C_2 and for unique solution of the specific problem with given initial values $f(0)$ and $f(1)$ it remains just to find the corresponding scalars C_1 and C_2. (And we also need to check whether it is possible for any two initial values).

3.11. Choice of scalars. Let us show how this can work on at least one example. Let us concentrate on the problem that the roots of the characteristic polynomial are in general not in the same field of the scalars as the coefficients in the equation. Thus we solve the problem:

$$\begin{aligned} y_{n+2} &= y_{n+1} + \frac{1}{2}y_n \\ y_0 &= 2,\ y_1 = 0. \end{aligned} \tag{3.5}$$

In our case is thus $\lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{3})$ 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 solutions for the equation with integral coefficients look complicated and are expressed with irrational (or possibly complex) numbers, we know a priori that the solution itself is again integral. Without this "step aside" into bigger field of scalars we would not be able to describe the general solution.

We will meet with similar events very often. General solution allows us also without direct enumeration of constant to discuss qualitative behaviour of the sequence of numbers $f(n)$, that is, whether the values with growing n approach some fixed value or oscillate in some interval or are unbounded.

3.12. General case of homogeneous recurrences. Let us now try similarly as in the case of second order to plug in the choice $x_n = \lambda^n$ for some (yet unknown) scalar λ into the general homogeneous equation from the definition 3.9. For every n we obtain the condition

$$\lambda^{n-k}(a_0\lambda^k + a_1\lambda^{k-1} \cdots + a_k) = 0$$

which means that either $\lambda = 0$ or λ is the root of the so-called *characteristic polynomial* in the parentheses. Characteristic polynomial is independent of n.

Assume that the characteristic polynomial has k distinct roots $\lambda_1, \dots, \lambda_k$. We can for this purpose extended the field of scalars we are working in, for instance $\mathbb{Q}$ into $\mathbb{R}$ or $\mathbb{R}$ into $\mathbb{C}$, because the result of the calculations will again solutions that stay in the original field thanks to the equation itself. Each of the roots gives us single possible solution

$$x_n = (\lambda_i)^n.$$

In order to be happy, we require k linearly independent solutions.

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 clever 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 out that $s = -\frac{1}{\sqrt{5}}$, $t = \frac{1}{\sqrt{5}}$ and thus

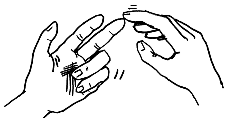
$$p_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n(\sqrt{5})}. \tag{3.2}$$

Such sequence satisfies the given recurrent formula and also the initial conditions $c_0 = 0$, $c_1 = 1$, thus it is the single sequence given by this requirements. Note that the value of the formula (‖3.2‖) is integer for any natural n (it gives the integer Fibonacci sequence), although it might not seem so at the first glance. □

3.5. Simplified model for behaviour of gross domestic product. Consider the difference equation

$$y_{k+2} - a(1+b)y_{k+1} + aby_k = 1, \tag{3.3}$$

where y_k is the gross domestic product at the year k. The constant a is the so-called *consumption tendency*, which is a macroeconomical factor that gives the fraction of money that the people spend (from what they have at their disposal), and the constant b describes the dependence of the measure of investment of the private sector on the consumption tendency.



We further assume that the size of the domestic product is normalised such that on the right-hand side of the equation the result is 1.

Compute the values y_n for $a = \frac{3}{4}$, $b = \frac{1}{3}$, $y_0 = 1$, $y_1 = 1$.

Solution. Let us first look 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 $a(\frac{1}{2})^n + bn(\frac{1}{2})^n$.

Let us also note that if we find some solution of the non-homogeneous equation (the so-called particular solution), then if we add to it any solution of the homogeneous solution, we obtain another solution of the non-homogeneous equation. It can be shown that by this we obtain all solutions of the non-homogeneous equation.

In our case (that is, when all the coefficients and the non-homogeneous term are constant) is a particular solution the constant

In order to do this it suffices to check the independence by plugging k values for $n = 0, \dots, k-1$ for k choices of λ_i into the Casortian (see 3.9). We thus obtain the so-called Vandermonde matrix and it is a nice (but not entirely trivial) exercise to compute that for every k and any k-tuple of distinct λ_i is determinant of such matrix non-zero, see ‖2.24‖ on the page 87. But that means that the chosen solutions are linearly independent.

We have thus found the fundamental system of solutions of the homogeneous difference equation in the case that all the roots of its characteristic polynomial are distinct.

Consider now the multiple root λ and plug into the definition the assumed solution $x_n = n\lambda^n$. We obtain the condition

$$a_0 n\lambda^n + \dots + a_k(n-k)\lambda^{n-k} = 0.$$

This condition can be rewritten using the so-called derivation of a polynomial (see **??** on the page **??**), which we denote by apostrophe:

$$\lambda(a_0\lambda^n + \dots + a_k\lambda^{n-k})' = 0$$

and right at the beginning of the fifth chapter we shall see that the root of a polynomial f has multiplicity greater than one if and only if it is a root of f'. Our condition is thus satisfied.

With greater multiplicity ℓ of the root of the characteristic polynomial we can proceed similarly and use the fact that a root with multiplicity ℓ is a root of all derivations of the polynomial up to $\ell - 1$ (inclusively). Derivations look like this:

$$\begin{aligned} f(\lambda) &= a_0\lambda^n + \dots + a_k\lambda^{n-k} \\ f'(\lambda) &= a_0 n\lambda^{n-1} + \dots + a_k(n-k)\lambda^{n-k-1} \\ f''(\lambda) &= a_0 n(n-1)\lambda^{n-2} + \dots + a_k(n-k)(n-k-1)\lambda^{n-k-2} \\ &\vdots \\ f^{(\ell+1)} &= a_0 n \dots (n-\ell)\lambda^{n-\ell-1} + \dots \\ &\qquad + a_k(n-k)\dots(n-k-\ell)\lambda^{n-k-\ell-1} \end{aligned}$$

Let us look on the case for a triple root λ and try to find a solution in the form $n^2\lambda^n$. Plugging into the definition we obtain the equation

$$a_0 n^2\lambda^n + \dots + a_k(n-k)^2\lambda^{n-k} = 0.$$

Clearly the left side equals the expression $\lambda^2 f''(\lambda) + \lambda f'(\lambda)$ and because λ is a root of both derivations, the condition is satisfied.

Using induction we easily prove that even for general condition for the solution in the form $x_n = n^\ell\lambda^n$,

$$a_0 n^\ell\lambda^n + \dots a_k(n-k)^\ell\lambda^{n-k} = 0,$$

the solution can be obtained as a linear combination of the derivations of the characteristic polynomial starting with the expression

$$\lambda^{\ell+1} f^{(\ell+1)} + \frac{1}{2}\lambda^\ell \ell(\ell+1) f^{(\ell)} + \dots$$

and we have thus came close to the complete proof of the following:

Theorem. *Every homogeneous linear difference equation of the order k over any field of scalars $\mathbb{K}$ contained in the complex numbers $\mathbb{K}$ has as a set of all solutions a 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*

$y_n = c$. By plugging into the equation we obtain $c - c + \frac{1}{4}c = 1$, that is, $c = 4$. 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 $a = b = -3$, 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 characterised by these properties. □

In the previous case we have used the so-called *method of indeterminate coefficients*. It is based on the following: on the basis of the non-homogeneous term of the given difference equation we "guess" the form of the particular solution. The forms of the particular solutions are known for many non-homogeneous terms. For instance the equation

$$y_{n+k} + a_1 y_{n+k-1} + \dots + a_k y_n = P_m(n), \tag{3.4}$$

where $P(m)$ is a polynomial of degree n and the corresponding characteristic equation has real roots, has (almost always) particular solution of the form $Q_m(n)$, where $Q_m(n)$ is a polynomial of degree m.

Other possible way how to solve such equation is the so-called *variation of constants method*, where we first find a solution

$$y(n) = \sum_{i=1}^{k} c_i f_i(n)$$

of the homogenised equation and we consider the constants c_i as functions $c_i(n)$ of the variable n and we look for a particular solution of the given equation in the form

$$y(n) = \sum_{i=1}^{k} c_i(n) f_i(n).$$

Let us show on the picture the values of f_i for $i \le 35$ with the equation

$$f(n) = \frac{9}{8} f(n-1) - \frac{3}{4} f(n-2) + \frac{1}{2}, \qquad f(0) = f(1) = 1.$$

natural numbers between zero and multiplicity of the corresponding root λ.

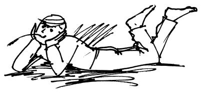
Proof. Aforementioned relations between the multiplicity of a root and the derivation of the polynomial will be proven later, and we won't prove the fact that every complex polynomial has exactly that many roots (counting multiplicities) as is its degree. Thus it just remains to prove that the found k-tuple of solutions is linearly independent. Even in this case we can inductively prove that the corresponding Casortian is non-zero, as we have already done in the case of Vandermonde determinant before.

For illustration of our approach we show how does the calculation look like for the case of a root λ_1 with multiplicity one and a root λ_2 with multiplicity two:

$$\begin{aligned}
C(\lambda_1^n, \lambda_2^n, n\lambda_2^n) &= \begin{vmatrix} \lambda_1^n & \lambda_2^n & n\lambda_2^n \\ \lambda_1^{n+1} & \lambda_2^{n+1} & (n+1)\lambda_2^{n+1} \\ \lambda_1^{n+2} & \lambda_2^{n+2} & (n+2)\lambda_2^{n+2} \end{vmatrix} \\
&= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 & \lambda_2 & (n+1)\lambda_2 \\ \lambda_1^2 & \lambda_2^2 & (n+2)\lambda_2^2 \end{vmatrix} \\
&= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 - \lambda_2 & 0 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & 0 & \lambda_2^2 \end{vmatrix} \\
&= -\lambda_1^n \lambda_2^{2n} \begin{vmatrix} \lambda_1 - \lambda_2 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & \lambda_2^2 \end{vmatrix} = \lambda_1^n \lambda_2^{2n+1} (\lambda_1 - \lambda_2)^2 \neq 0.
\end{aligned}$$

In the general case the proof can be carried on in a completely similar way, inductively. □

3.13. Real basis of the solutions. For equations with real coefficients the initial real conditions always lead to real solutions. Still, the corresponding fundamental solutions derived using the just proven theorem might exists only in the complex domain.



Let us therefore try to find other generators, which will be more convenient for us. Because the coefficients of the characteristic polynomial are real, each its root is either real or the roots are paired as complex conjugates.

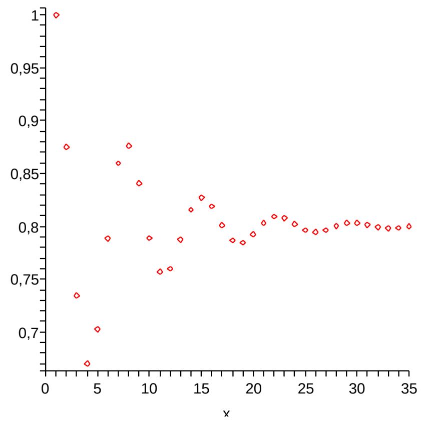
If we describe the solution in the polar form as

$$\begin{aligned}
\lambda^n &= |\lambda|^n (\cos n\varphi + i \sin n\varphi) \\
\bar{\lambda}^n &= |\lambda|^n (\cos n\varphi - i \sin n\varphi),
\end{aligned}$$

we immediately see that their sum and difference leads to other 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. Nice topic to think about is connection between absolute values of individual roots and stability of the solution – either of all of them or with dependence on the initial conditions. We will not go into details here, because only in the fifth chapter we will speak of convergence of values to some limit value and so on, but still there is some space fore some interesting numerical experiments for instance with oscillations of suitable population or economical models.



Let us 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 give some two neighbouring members. Let us again note a further usability of complex numbers: for determining the explicit formula for the n-th member of the sequence of real numbers we might require calculations with complex numbers (that happens when the characteristic polynomial of the difference equation has complex roots).

3.6. Find 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. Homogenised equation is

$$x_{n+2} = 2x_n.$$

Its characteristic polynomial is $x^2 - 2$, its roots are $\pm\sqrt{2}$. The solution of the homogenised 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, particular solution will thus be in the form of 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 at 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 $-n - 2$.

3.14. Non-homogeneous linear difference equations.

Analogously to 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, and $a_0(n) \neq, a_k(n) \neq 0$.

We proceed by finding one solution and adding whole vector space of dimension k of solutions of the corresponding homogeneous system. Indeed this yields a solution and because the 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 usually not easy to find that one particular solution of a non-homogeneous system, if the behaviour of the scalar coefficients in the equation is complicated. For linear recurrences the situation is similar.

Let us restrict ourselves to single case, where 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, $i = 1, \dots, s$. By plugging into the difference equation and comparing the coefficients at 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, we have found a solution of our original problem. If it has no solution, it is enough to increase the degree s of the polynomial in question.

For instance the equation $x_n - x_{n-2} = 2$ cannot have a constant solution, but by setting $x_n = \alpha_0 + \alpha_1 n$ we obtain the solution $\alpha_1 = 1$ (and the coefficient α_0 can be arbitrary) and thus the general solution of our equation is

$$x_n = C_1 + C_2(-1)^n + n.$$

Note that the matrix of the corresponding system of equations for a polynomial of lower degree is zero and the equation $0 \cdot \alpha_0 = 2$ has no solution.

3.15. Linear filters. We will now consider now the infinite sequences

$$x = (\dots, x_{-n}, x_{-n+1}, \dots, x_{-1}, x_0, x_1, \dots, x_n, \dots)$$

and will work, similarly to the case of systems of linear equations, with the operation T that maps the whole sequence x to the sequence $z = Tx$ with elements

$$z_n = a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k}.$$

With sequences x we can again work as with vectors, operations working coordinate-wise, and the vector space is infinitely-dimensional. Our mapping T is clearly a linear mapping in such space.

The sequences can be imagined as discrete values of some signal, subtracted usually in very short time units, operation T can then be a filter that works with the signal. We are interested in estimating the properties such "filter" can have.

Thus the solution of the non-homogeneous difference equation of the second order without initial condition is of the form $a(\sqrt{2})^n + b(-\sqrt{2})^n - n - 2$, $a, b \in \mathbb{R}$.

Now, by plugging in the initial conditions, we determine the indeterminate $a, b \in \mathbb{R}$. For calculational simplicity 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:

$$\begin{aligned} x_0 : &\quad a(\sqrt{2})^0 + b(-\sqrt{2})^0 - 2 = 1, \quad \text{thus } a + b = 3,\\ x_1 : &\quad \sqrt{2}a - \sqrt{2}b = 5, \end{aligned}$$

whose solution gives us $a = \frac{6+5\sqrt{2}}{4}$, $b = \frac{6-5\sqrt{2}}{4}$. The solution is thus the sequence

$$x_n = \frac{6+5\sqrt{2}}{4}(\sqrt{2})^n + \frac{6-5\sqrt{2}}{4}(-\sqrt{2})^n - n - 2.$$

□

3.7. Determine the real basis of the space of all solutions of the homogeneous difference equation

$$x_{n+4} = x_{n+3} + x_{n+1} - x_n,$$

Solution. The characteristic polynomial of the given equation is $x^4 - x^3 - x + 1$. If we are looking for its roots, we are solving the reciprocal equation

$$x^4 - x^3 - x + 1 = 0$$

Standard procedure is to solve the equation by the expression x^2 and then we use the substitution $t = x + \frac{1}{x}$, that is, $t^2 = x^2 + \frac{1}{x^2} + 2$. We obtain the equation

$$t^2 - t - 2 = 0,$$

with roots $t_1 = -1$, $t_2 = 2$. For both of these values of the indeterminate t we solve separately the equation given by the substitution:

$$x + \frac{1}{x} = -1.$$

It has two complex roots: $x_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = \cos(2\pi/3) + i\sin(2\pi/3)$ and $x_2 = -\frac{1}{2} - i\frac{\sqrt{3}}{2} = \cos(2\pi/3) - i\sin(2\pi/3)$.

For the second value of the indeterminate t we obtain the equation

$$x + \frac{1}{x} = 2$$

Signals are very often from their essence a sum of some parts, which are by themselves basically 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 $z = Tx$

$$\begin{aligned} z_{n+p} &= a_0x_{n+p} + a_1x_{n-1+p} + \cdots + a_kx_{n-k+p}\\ &= a_0x_n + a_1x_{n-1} + \cdots + a_kx_{n-k} = z_n \end{aligned}$$

with the same period p.

For a fixed operation T we are interested in the following: which input periodic sequences remain roughly the same (up to a multiple) and which will be suppressed to zero values.

In the second case we are looking for the kernel of our linear mapping T. That is given by homogeneous difference equation

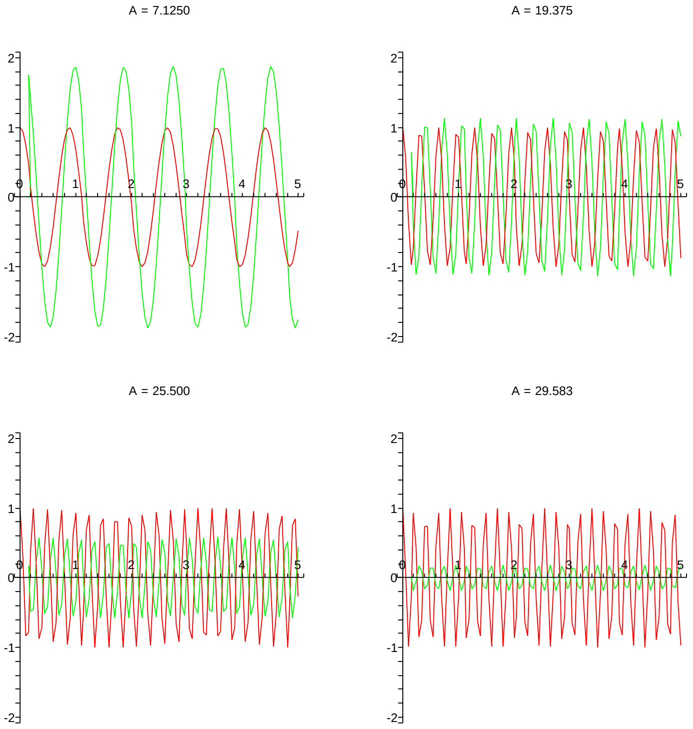
$$a_0x_n + a_1x_{n-1} + \cdots + a_kx_{n-k} = 0, \quad a_0 \neq 0 \quad a_k \neq 0,$$

which we are able to solve.

3.16. Bad equaliser. As an example, let us consider a very simple linear filter given by the equation

$$z_n = (Tx)_n = x_{n+2} + x_n.$$

The results of such operation on a signal are hinted at in the following four pictures for gradually increasing frequency of periodic signal $x_n = \cos(\varphi n)$. Red is the original signal, green is the result after using the filter. Unevenness of the curves are consequence of imprecise depicting, both signals are of course smooth sinus curves.



Note that in the areas where the resulting signal is roughly as strong as the original there is a dramatic shift in the phase. Cheap equaliser indeed work in such a bad way.

with 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: $\{\left(-\frac{1}{2}+i\sqrt{3}\right)^n\}_{n=1}^{\infty}$, $\{\left(-\frac{1}{2}-i\sqrt{3}\right)^n\}_{n=1}^{\infty}$, $\{1\}_{n=1}^{\infty}$ (constant sequence) and $\{n\}_{n=1}^{\infty}$. If we are looking for a real basis, 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 basis of the solution space: $\{1\}_{n=1}^{\infty}$ (constant sequence), $\{n\}_{n=1}^{\infty}$, $\{\cos(n\cdot 2\pi/3)\}_{n=1}^{\infty}$, $\{\sin(n\cdot 2\pi/3)\}_{n=1}^{\infty}$. □

3.8. Find a sequence that 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. General solution of the homogenised equation is of the form $a(-1)^n + b2^n$. Particular solution is the constant $-1/2$. General solution of the given non-homogeneous equation without initial conditions is thus

$$a(-1)^n + b2^n - \frac{1}{2}.$$

By plugging in the initial conditions then gives us the constants $a = -5/6$, $b = 5/6$. The given difference equation with initial conditions is thus satisfied by the sequence

$$-\frac{5}{6}(-1)^n + \frac{5}{3}2^{n-1} - \frac{1}{2}.$$

□

3.9. Determine the sequence of real numbers that satisfies the following non-homogeneous difference equation with initial conditions:

$$2x_{n+2} = -x_{n+1} + x_n + 2, \quad x_1 = 2,\ x_2 = 3.$$

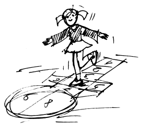
Solution. General solution of the homogenised equation is of the form $a(-1)^n + b(1/2)^n$. Particular solution is the constant 1. General solution of the non-homogeneous equation without initial conditions is thus

$$a(-1)^n + b\left(\frac{1}{2}\right)^n + 1.$$

By plugging into the initial conditions we obtain the constants $a = 1$, $b = 4$. The given equation with initial conditions is thus satisfied by the sequence

$$(-1)^n + 4\left(\frac{1}{2}\right)^n + 1.$$

3. Iterated linear processes



3.17. Iterated processes. In practical models we very 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 iteration. Very often is the linear process remains the same, from the mathematical point of view it is thus repeated multiplication of the state vector by the same matrix.

While for solving the systems of linear equation we needed only minimal knowledge of properties of linear mappings, in order to understand the behaviour of an iterated system we need to know the properties of eigenvalues, properties of eigenvectors and other structural results.

In a sense we are in the same environment as with linear recurrences and actual our description of filters in previous paragraphs can be described in such way. Imagine that we are working with sound and are keeping track by the state vector

$$Y_n = (x_n, \dots, x_{n-k+1})$$

of all values from the actual one to the last one that is yet being processed in our linear filter. In one time interval (for the frequency of audio signal a very short one) we then move to the state vector

$$Y_{n+1} = (x_{n+1}, x_n, \dots, x_{n-k+2}),$$

where the first value $x_{n+1} = a_1x_n + \dots + a_kx_{n-k+1}$ is computed as with homogeneous difference equations, the others are just shift by one position and last one is forgotten. The corresponding square matrix of order k that satisfies $Y_{n+1} = A\cdot Y_n$ looks as follows:

$$A = \begin{pmatrix} a_1 & a_2 & \dots & a_{k-1} & a_k \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0. \end{pmatrix}$$

For such simple matrix we have derived explicit procedure for the complete formula for the solution. In general, it wont be so easy even for very similar systems. One of the typical cases is study of dynamics of a population in distinct biological systems.

Note also that the matrix A has (understandably) the characteristic polynomial

$$p(\lambda) = \lambda^k - a_1\lambda^{k-1} - \dots - a_k,$$

as can be easily derived by expanding the last column and the recurrence. That is explainable also directly, because the solution $x_n = \lambda^n$, $\lambda \neq 0$ basically means that the matrix A by multiplication takes the eigenvector $(\lambda^k, \dots, \lambda)^T$ to its λ-multiple. Thus such λ must be eigenvalue of the matrix A.

3.18. 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. Linear model of evolution of such system is then given by the matrix A of dimension n, which gives the change of the vector X_n

□

3.10. 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, $k = 1, \dots, n$, are equal). Thus we 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

$$x_{1,2,3,4} = \frac{1 \pm \sqrt{5} \pm \sqrt{-10 \pm 2\sqrt{5}}}{4}.$$

Now we note that the roots of the characteristic equation could have been "guessed" right away – it is

$$x^5 + 1 = (x+1)(x^4 - x^3 + x^2 - x + 1),$$

and 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 by the way 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 $\sin(\frac{3\pi}{5}) = \frac{\sqrt{10+2\sqrt{5}}}{4}$ (because all the roots of the equation have the absolute value 1, they are real (imaginary) parts of the corresponding roots). □

3.11. Determine the explicit expression of the sequence satisfying the difference equation $x_{n+2} = 2x_{n+1} - 2x_n$ with members $x_1 = 2$, $x_2 = 2$.

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

to

$$X_{n+1} = A \cdot X_n$$

when time changes from t_n to t_{n+1}.

Let us show as an example the so-called *Leslie model for population growth*, where 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 relative mortality in the i-th group in one time interval. Clearly such 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 more general theory, let us play for a while with 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:

$$p_m(\lambda) = |A - \lambda E| = -\lambda p_{m-1}(\lambda) + (-1)^{m-1} f_m \tau_1 \dots \tau_{m-1}.$$

Easily by induction we derive that this characteristic polynomial is of the form

$$p_m(\lambda) = (-1)^m (\lambda^m - a_1 \lambda^{m-1} - \dots - a_{m-1}\lambda - a_m)$$

and mainly non-negative coefficients $a_1, \dots, a_m$, if all parameters τ_i and f_i are positive. For instance it is always

$$a_m = f_m \tau_1 \dots \tau_{m-1}.$$

Let us qualitatively estimate the distribution of the roots of the polynomial p_m. Sadly, details of this procedure could be exactly explained only later, after understanding some parts of the so-called mathematical analysis in the chapter five and later, however it should all be intuitively clear even now. We express 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 $\lambda > 0$. 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 there exists exactly one positive real eigenvalue.

For actual Leslie models of populations all coefficients τ_i and f_j are between one and zero and 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 smaller than one.

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

$$\begin{aligned} 1 = x_0 &= a_1 + ia_2 + b_1 + ib_2 \\ 2 = x_1 &= (a_1 + ia_2)(1+i) + (b_1 + ib_2)(1-i), \end{aligned}$$

and by comparing the real and the complex part of both equations we obtain a linear system of four equations with four indeterminates

$$\begin{aligned} a_1 + b_1 &= 1 \\ a_2 + b_2 &= 0 \\ a_1 - a_2 + b_1 + b_2 &= 2 \\ a_1 + a_2 - b_1 + b_2 &= 0 \end{aligned}$$

with solution $a_1 = b_1 = b_2 = \frac{1}{2}$ and $a_2 = -1/2$. Thus we can express the sequence in question as

$$x_n = (\frac{1}{2} - \frac{1}{2}i)(1+i)^n + (\frac{1}{2} + \frac{1}{2}i)(1-i)^n.$$

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

$$\begin{pmatrix} c \\ d \end{pmatrix} = T^{-1} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

thus we have again an alternative expression of the sequence x_n where there are no complex numbers (but there are square roots):

$$x_n = (\sqrt{2})^n \cos\left(\frac{n\pi}{4}\right) + (\sqrt{2})^n \sin\left(\frac{n\pi}{4}\right),$$

which we could have obtained 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 $2 = x_1 = c \cdot u_1 + d \cdot v_1 = c + d$. □

If we begin with any state vector X which is given as a sum of eigenvectors

$$X = X_1 + \cdots + X_m$$

with eigenvalues λ_i, then iterations yield

$$A^k \cdot X = \lambda_1^k X_1 + \ldots \lambda_m^k X_m,$$

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 $\lambda_1 X_1^k$.

Distribution of the population among the age groups are thus very fast approaching the ratios of the components of eigenvector to the dominant eigenvalue λ_1.

For example for the matrix (let us realise the meaning of individual coefficient, they 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 immediately chosen the eigenvector whose coordinates sum to 100, it directly gives us the percentual distribution of the population.

If we instead of three-percent total growth of the population rather wanted constant number and said that we will eat sheep from second group, we would be asking the question how much shall we decrease τ_2 so that the dominant eigenvalue would be one.

3.19. Matrices with non-negative elements. Real matrices that have no negative elements have very special properties. Also, they are very often present in practical models. We shall thus introduce the so-called *Perron-Frobenius theory* which deals with such matrices.

Let us begin with definition of some notions in order to be able to formulate our ideas.

Positive and primitive matrix

Definition. Under *positive matrix* we understand a square matrix A whose all elements a_{ij} are real and strictly positive. *Primitive matrix* is such square matrix A such that some power A^k is positive.

Let us recall that *spectral radius of matrix A* is the maximum of absolute values of all (complex) eigenvalues of A. Spectral radius of a linear mapping over (finitely-dimensional) vector space is the spectral radius of the corresponding matrix under some basis.

Norm of a matrix $A \in \mathbb{R}^{n^2}$ or of a vector $x \in \mathbb{R}^n$ is the sum of absolute values of all elements. For vector x we write $|x|$ for its norm.

The following result is very useful and hopefully also well understandable. Its proof is with its hardness quite atypical for this

3.12. Determine the explicit expression of the sequence satisfying the difference equation $x_{n+2} = 3x_{n+1} + 3x_n$ with members $x_1 = 1$ and $x_2 = 3$. ○

3.13. Determine the explicit formula for the n-th member of the unique solution $\{x_n\}_{n=1}^{\infty}$ that satisfies the following conditions:

$$x_{n+2} = x_{n+1} - x_n, \ , x_1 = 1, \ x_2 = 5.$$

○

3.14. Determine the explicit formula for the n-th member of the unique solution $\{x_n\}_{n=1}^{\infty}$ that satisfies the following conditions:

$$-x_{n+3} = 2x_{n+2} + 2x_{n+1} + x_n, \ x_1 = 1, \ x_2 = 1, \ x_3 = 1.$$

○

3.15. Determine the 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 are to deal with right now will 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.

3.16. Savings. With a friend we are saving 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 many will I pay in the twelfth month?

Solution. The amount of many I pay in the n-th month is denoted as x_n and the amount my friend is paying is y_n. The first month we thus give $x_1 = 10$, $y_1 = 20$. For the following payments we can write down a recurrent relation:

$$\begin{aligned} x_{n+1} &= x_n + \tfrac{1}{2} y_n \\ y_{n+1} &= y_n + \tfrac{1}{2} x_n \end{aligned}$$

If we denote the common savings as $z_n = x_n + y_n$, then by summing the equations we obtain $z_{n+1} = z_n + \frac{1}{2} z_n = \frac{3}{2} z_n$. That is a geometric sequence and we obtain $z_n = 3.(\frac{3}{2})^{n-1}$. In a year we will thus have $z_1 + z_2 + \cdots + z_{12}$. This partial sum is easy to compute

$$3(1 + \frac{3}{2} + \cdots + (\frac{3}{2})^{11}) = 3\frac{(\frac{3}{2})^{12} - 1}{\frac{3}{2} - 1} \doteq 772,5.$$

textbook, so we give at least a vague idea how to do it. If the reader has some problems with smooth reading, we suggest skipping the proof immediately.

Theorem (Perron)**.** *If A is a primitive matrix with spectral radius $\lambda \in \mathbb{R}$, then λ is a simple root of characteristic polynomial of the matrix A, which is strictly greater than the absolute value of any other eigenvalue of the matrix A. Furthermore, there exist eigenvector x associated with λ such that all elements x_i of x are positive.*

Vague idea. In the proof we shall rely on the intuition from elementary geometry. Partly we will make the used concepts more precise in the analytical geometry in the fourth chapter, some analytical aspects will be studied in more detail in the fifth chapter and later, and some claims won't be proven in this textbook at all. Hopefully the presented ideas will not just illuminate the theorem but also will motivate for deeper study of geometry and analysis by themselves. Let us begin with a understandable auxiliary lemma:

Lemma. *Consider any polyhedron P containing the origin $0 \in \mathbb{R}^n$. If some iteration of the linear mapping $\psi : \mathbb{R}^n \to \mathbb{R}^n$ maps P in its inside, then the spectral radius of the mapping ψ is strictly smaller than one.*

Consider the matrix A of the mapping ψ under the standard basis. Because the eigenvalues of A^k are the k-th powers of the eigenvalues of the matrix A, we can without loss of generality assume that the mapping ψ already maps P into its inside. Clearly ψ cannot have any eigenvalue with absolute value greater than one.

Let us argue by contradiction. Assume that there exists eigenvalue λ with $|\lambda| = 1$. Thus 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 when the points in the image group about some point y in $\mathbb{R}^n$, the point y is also in the image) and the border of P is not intersected at all by the image. Thus ψ cannot have a fixed point on the border and there cannot even be any point on the border to which the points in the image would converge. The first argument excludes that some power of λ is one, because such fixed point on the border of P would then exist. In the remaining case there would definitely be a two-dimensional subspace $W \subset \mathbb{R}^n$ on which the restriction of ψ acts as a rotation by an irrational argument and thus there definitely exist a point y in the intersection of W with the border of P. But then the point y could be approached arbitrarily close by the points from the set $\psi^n(y)$ (through all iterations) and thus would have to be in the image also. That leads to a contradiction and thus the lemma is proven.

Now let us prove the Perron theorem. Our first step is ensuring the existence of the eigenvector which has all elements positive. Let us consider the so-called standard simplex

$$S = \{x = (x_1, \dots, x_n)^T, \ |x| = 1, x_i \geq 0, i = 1, \dots, n\}.$$

Because all elements in the matrix A are non-negative, the image $A \cdot x$ has all non-negative coordinates as x does and at least one of them is always non-zero. 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 so-called Brouwer fixed point theorem and thus there exists vector $y \in S$ such that it is mapped by this mapping to itself. That

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 $(1-\lambda)^2 - \frac{1}{4} - 0$ and the eigenvalues are thus $\lambda_{1,2} = \frac{3}{2}, \frac{1}{2}$. 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 12. month I pay

$$x_{12} = \left(\frac{3}{2}\right)^{12} - \left(\frac{1}{2}\right)^{12} \doteq 130$$

EUR and my friend pays basically the same amount. □

Remark. The previous example can be solved also without matrices by rewriting the recurrent equation: $x_n = x_n + \frac{1}{2} y_n = \frac{1}{2} x_n + \frac{1}{2} z_n$.

The previous example was actually a model of growth (in the case of growth of saved money). Let us now go to the models of growth describing primarily a growth of some population. Leslie model of population growth with which we have coped 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:

3.17. Rabbits for the second time. Let us show how the Leslie model can describe the population of the rabbits on the meadow with which we have worked in the exercise (‖3.4‖). Let us consider that the rabbits are dying after reaching the ninth year of age (in the original model the rabbits were immortal). Let us denote the numbers of rabbits according to their age in months in time t (in months) as $x_1(t)$, $x_2(t)$,..., $x_9(t)$, then the numbers of rabbits in individual categories are after one month described by the formula $x_1(t+1) =$

means that

$$A \cdot y = \lambda\, y, \quad \lambda = |A \cdot y|$$

and we have found an eigenvector that lies in S. Because some power of A^k has due to our assumption all elements positive and of course we have $A^k \cdot y = \lambda^k y$, all elements of the vector y are strictly positive (that is, they lie inside of S) and $\lambda > 0$.

In order to prove the rest of the theorem, we will consider the mapping given by the matrix A in a more suitable basis and furthermore we shall multiply it by a constant λ^{-1}:

$$B = \lambda^{-1}(Y^{-1} \cdot A \cdot Y),$$

where Y is a diagonal matrix with coordinates y_i of a just-found eigenvector y on a diagonal. Evidently B is also a primitive matrix and furthermore the vector $z = (1, \dots, 1)^T$ is its eigenvector, because clearly $Y \cdot z = y$.

If we know 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, the proof is finished.

In order to do that we use the auxiliary lemma. Consider the matrix B to be a matrix of a linear mapping that maps the row vectors

$$u = (u_1, \dots, u_n) \mapsto u \cdot B = v,$$

that is, using multiplication from the right. Thanks to the fact that $z = (1, \dots, 1)^T$ is an eigenvector of the matrix B, the sum of the coordinates of the row vector v

$$\sum_{i,j=1}^{n} u_i b_{ij} = \sum_{i=1}^{n} u_i = 1,$$

whenever $u \in S$. Therefore the mapping maps the simplex S on itself and thus has in S a (row) eigenvector w with eigenvalue one (fixed point, thanks to the theorem of Brouwer). Because some power B^k contains only strictly positive elements, the image of the simplex S in the k-th iteration of the mapping given by B lies inside of S. We are getting close to using our lemma prepared for this proof.

We shall still work 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 polyhedron containing the origin and the vector subspace $V \subset \mathbb{R}^n$ generated by P is invariant to the action of the matrix B through multiplication of the row vectors from the right. Restriction of our mapping on P thus satisfies the assumptions of the auxiliary lemma and thus all its eigenvalues are strictly smaller than one.

We have yet to deal with the problem that the just considered mapping is given by multiplication of the row vectors from the right with the matrix B, while originally we were interested in the mapping given by the matrix B and multiplication of the column vectors from the left. But that 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.

Dimension of the space V is $n-1$, thus completing the proof. □

$x_2(t) + x_3(t) + \cdots + x_9(t)$, $x_i(t+1) = x_{i-1}(t)$, pro $i = 2, 3, \ldots, 10$, or

$$\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ x_4(t+1) \\ x_5(t+1) \\ x_6(t+1) \\ x_7(t+1) \\ x_8(t+1) \\ x_9(t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \\ x_7(t) \\ x_8(t) \\ x_9(t) \end{pmatrix}.$$

Characteristic polynomial of the given matrix is $\lambda^9 - \lambda^7 - \lambda^6 - \lambda^5 - \lambda^4 - \lambda^3 - \lambda^2 - \lambda - 1$. Roots of this equation 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 $1,608^t$.

3.18. Pond. Let us have a simple model of a pond where there lives a population of white fish (roach, bleak, vimba, nase, etc.). We 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 survives 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 population would clearly very quickly fill the pond. Thus we want to maintain a balance by using a 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 stagnate?

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, 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

$$\begin{gathered} \tau - 1 + 0.36 - 0.6(\tau - 1) = 0 \\ 0.4\tau - 0.04 = 0 \end{gathered}$$

3.20. Simple corollaries. The following very useful claim has with the knowledge of the Perron theorem a surprisingly simple proof and shows how strong is the property of the primitive matrix of a mapping.

Corollary. *If $A = (a_{ij})$ is a primitive matrix and $x \in \mathbb{R}^n$ its eigenvector with all coordinates non-negative and eigenvalue λ, then $\lambda > 0$ is the spectral radius of A. Furthermore it holds that*

$$\min_{j \in \{1,\ldots,n\}} \sum_{i=1}^{n} a_{ij} \le \lambda \le \max_{j \in \{1,\ldots,n\}} \sum_{i=1}^{n} a_{ij}.$$

Proof. Consider the eigenvector x from the statement. Because A is primitive, we can fix k such that A^k has only positive elements, then of course $A^k \cdot x = \lambda^k x$ is a vector with all coordinates strictly positive. Necessarily then $\lambda > 0$.

>From the theorem of Perron we know that the spectral radius μ is an eigenvalue and choose such eigenvector y associated with μ such that the difference $x - y$ has only strictly positive coordinates. Then necessarily for all the powers of n we have

$$0 < A^n \cdot (x - y) = \lambda^n x - \mu^n y,$$

but we also have that $\lambda \le \mu$. From there we directly have $\lambda = \mu$.

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 a vector with the sum of coordinates equal to one and count:

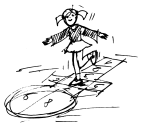
$$\begin{gathered} \sum_{i,j=1}^{n} a_{ij} x_j = \sum_{i=1}^{n} \lambda x_i = \lambda \\ \lambda = \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} a_{ij} \Big) x_j \le \sum_{j=1}^{n} b_{\max} x_j = b_{\max} \\ \lambda = \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} a_{ij} \Big) x_j \ge \sum_{j=1}^{n} b_{\min} x_j = b_{\min}. \end{gathered}$$

□

Note that for instance all Leslie matrices from 3.18, where all the coefficients f_i and τ_j are strictly positive, are primitive and thus we can apply on them the just derived results.

Perron-Frobenius theorem is a generalisation of the Perron theorem for more general matrices, which we won't give here. More information can be found for instance in **??**.

3.21. Markov chains. Very frequent and interesting case of linear processes with only non-negative elements in matrix is a mathematical model of a system which can be in one of m states with various probabilities. In a given point of time the system is in state i with probability x_i and 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 probability vector

$$x_n = (u_1(n), \ldots, u_m(n))^T.$$

That means that all components of the vector x are real non-negative numbers and their sum equals one. Components give the distribution of the probability of individual possibilities for the state of the system. The distribution of the probabilities at time

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 should according to the previous computation 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 approximately one esox for 556 white fish. □

In general, we can work with the previous model as follows:

3.19. Let in the population model prey-predator be the relation between the number of predators D_k and preys K_k in the given and the following month ($k \in \mathbb{N} \cup \{0\}$) be given by the linear system

(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)
$$\begin{aligned} D_{k+1} &= 0.6\,D_k + 0.5\,K_k, \\ K_{k+1} &= -0.135\,D_k + 1.2\,K_k. \end{aligned}$$

Let us analyse the behaviour of this model after a very long time.

Solution. Note that individual variants differ from each other only in the value of the coefficients at D_k in the second equation. We can thus 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 gradually 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 (clearly a „humble“) predator per month. When denoting

$$T = \begin{pmatrix} 0.6 & 0.5 \\ -a & 1.2 \end{pmatrix}$$

we immediately 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 preys after a very long time.

We easily compute the eigenvalues

(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 matrix T is and the respective eigenvectors are

(a) $(5,4)^T, \quad (5,2)^T;$

(b) $(10,7)^T, \quad (2,1)^T;$

(c) $(10,9)^T, \quad (10,3)^T.$

$n+1$ is given by multiplying the probabilistic transition matrix $T = (t_{ij})$, that is,

$$x_{n+1} = T \cdot x_n.$$

Because we assume that the vector x captures all possible states and thus with total probability one again transits into some of the state, all columns of T are also given by probabilistic vectors. Such process is called (discrete) *Markov process* and the resulting sequence of vectors $x_0, x_1, \dots$ is called *Markov chain* x_n.

Note that every probabilistic vector x is actually mapped by a Markov process on a vector with a sum of coordinates equal to one:

$$\sum_{i,j} t_{ij} x_j = \sum_j \Big(\sum_i t_{ij} \Big) x_j = \sum_j x_j = 1.$$

Now we can use 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 can easily see that the matrix $T - E$ is singular and thus one is surely an eigenvalue of the matrix T.

If furthermore T is a primitive matrix (for instance, when all elements are non-zero), from the corollary 3.20 we know that one is a simple root of the characteristic polynomial and all others have absolute value strictly smaller than one.

Theorem. *Markov processes with the matrix that has no zero element or that some its power has this property, satisfy:*

- *there exist unique eigenvector x_∞ for the eigenvalue* 1 *which is probabilistic,*
- *the iteration $T^k x_0$ approaches the vector x_∞ for any initial probabilistic vector x_0.*

Proof. This claim follows directly from the positivity of the coordinates of the eigenvector derived in the Perron theorem.

Assume first that the algebraic and geometric multiplicities of the eigenvalues of the matrix T are the same. Then every probabilistic vector x_0 can be (in complex extension $\mathbb{C}^n$) written as linear combination

$$x_0 = c_1 x_\infty + c_2 u_2 + \dots + c_n u_n,$$

where $u_2 \dots, u_n$ extend x_∞ to a basis of the eigenvectors. But then the k-th iteration gives again a probabilistic vector

$$x_k = T^k \cdot x_0 = c_1 x_\infty + \lambda_2^k c_2 u_2 + \dots + \lambda_n^k c_n u_n.$$

Because all eigenvalues $\lambda_2, \cdots \lambda_n$ are in absolute value strictly smaller than one, all components of the vector x_k but the first one approach (in norm) zero very rapidly. But x_k is still probabilistic, thus it must be that $c_1 = 1$ and the second claim is proven.

In reality even with distinct algebraic and geometric multiplicities of eigenvalues we reach the same conclusion using a more detailed study of the so-called root subspaces of the matrix T which we reach in the connection with the so-called Jordan matrix decomposition even in this chapter, see the note 3.33.

Even in the general case we reach in the eigensubspace $\langle x_\infty \rangle$ a uniquely determined invariant $(n-1)$-dimensional complement, on which are all eigenvalues in absolute value smaller than one and thus the corresponding component in x_k approaches zero as before. □

For $k \in \mathbb{N}$ thus holds that

(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 further have for big $k \in \mathbb{N}$ that

(a)

$$\begin{aligned} T^k &\approx \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix}^{-1} \\ &= \frac{1}{10} \begin{pmatrix} -10 & 25 \\ -8 & 20 \end{pmatrix}; \end{aligned}$$

(b)

$$\begin{aligned} T^k &\approx \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}; \end{aligned}$$

(c)

$$\begin{aligned} T^k &\approx \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1,05^k & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 10 & 10 \\ 9 & 3 \end{pmatrix}^{-1} \\ &= \frac{1,05^k}{60} \begin{pmatrix} -30 & 100 \\ -27 & 90 \end{pmatrix}, \end{aligned}$$

because exactly for big $k \in \mathbb{N}$ 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)

$$\begin{pmatrix} 1.05 & 0 \\ 0 & 0.75 \end{pmatrix}^k \approx \begin{pmatrix} 1.05^k & 0 \\ 0 & 0 \end{pmatrix}.$$

Let us note that in the variant (b), that is for $a = 0.175$, it was not necessary to compute the eigenvectors.

Thus we have obtained

(a)

$$\begin{aligned} \begin{pmatrix} D_k \\ K_k \end{pmatrix} &\approx \frac{1}{10} \begin{pmatrix} -10 & 25 \\ -8 & 20 \end{pmatrix} \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix} \\ &= \frac{1}{10} \begin{pmatrix} 5\,(-2D_0 + 5K_0) \\ 4\,(-2D_0 + 5K_0) \end{pmatrix}; \end{aligned}$$

3.22. Iteration of the stochastic matrices. Matrices of Markov chains, that is, matrices whose rows have sum of their components equal to one are called *stochastic matrices*. Standard problems connected with Markov processes contain answers to the question about the expected time elapsed between transition from one state to another and so on. Right now we are not prepared for solving these problems, but we return to this topic later.

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 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 theorem before). 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 probabilistic vectors and thus all of them converge to x_∞. □

Now for a short goodbye to Markov processes we think about the problem whether there exist for a given system the states into which the system tends to get in and stay in them.

We say that a state is *transient*, if the system stays in it with probability strictly smaller than one. State is *absorbing* if the system stays in it with probability one and into which the system can get with non-zero probability from any of the transient states. Finally, Markov chain x_n is *absorbing*, if all its states are either absorbing or transient.

If in the absorbing Markov chain first of r states of the system are absorbing, then for the stochastic matrix T of the system this means that it decomposes into "block-wise" upper triangular form

$$T = \begin{pmatrix} E & R \\ 0 & Q \end{pmatrix}$$

where E is a unit matrix whose dimension is given by the number of absorbing states, while R is a positive matrix and Q nonnegative. In any case, iterations of this matrix yield a matrix which has the same block of zero values in the bottom-left block and thus it is not primitive, for instance

$$T^2 = \begin{pmatrix} E & R + R \cdot Q \\ 0 & Q^2 \end{pmatrix}.$$

Even about such matrices we can obtain many information using the full Perron-Frobenius theory and with knowledge of probability and statistics also estimate expected time after which the system gets into one of the absorbing states.

4. More matrix calculus

On pretty practical examples we have seen that understanding the inner structure of matrices and their properties is a strong tool for specific computations and analyses. Even more is it true for effectivity of numerical calculations with matrices. Therefore we will for a while deal with abstract theory.

(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)

$$\begin{aligned} \begin{pmatrix} D_k \\ K_k \end{pmatrix} &\approx \frac{1.05^k}{60} \begin{pmatrix} -30 & 100 \\ -27 & 90 \end{pmatrix} \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix} \\ &= \frac{1.05^k}{60} \begin{pmatrix} 10\,(-3D_0 + 10K_0) \\ 9\,(-3D_0 + 10K_0) \end{pmatrix}. \end{aligned}$$

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.

Even a tiny change of the size of a can lead to a completely different result. This is caused by the constantness 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. □

3.20. Remark. Other 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.

In linear models an important role is played by the primitive matrices (3.19).

3.21. Which of the matrices

$$A = \begin{pmatrix} 0 & 1/7 \\ 1 & 6/7 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 1/6 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 0 & 1/2 \\ 3/4 & 0 & 1/2 \end{pmatrix},$$

$$D = \begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/6 & 1/6 & 1/3 \\ 1/6 & 0 & 5/6 & 2/3 \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

are primitive?

Solution. Because

$$A^2 = \begin{pmatrix} 1/7 & 6/49 \\ 6/7 & 43/49 \end{pmatrix}, \quad C^3 = \begin{pmatrix} 3/8 & 1/4 & 1/4 \\ 1/4 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/2 \end{pmatrix},$$

We will investigate further some special types of linear mappings on vector spaces and also a general case where the structure is described using the so-called Jordan theorem.

3.23. Unitary spaces and mappings. We are already used to the fact that it is efficient to work in the domain of complex numbers even in the case when we are interested only in real objects. Furthermore, in many areas the complex vector spaces are necessary component of the problem. For instance, take the so-called quantum computing, which became a very active area of theoretical computer science, although quantum computers have not been constructed yet (in a usable form).

Therefore we extend what we know about orthogonal mappings and mappings from the end of the second chapter with the following definitions:

Unitary spaces

Definition. *Unitary space* is a complex vector space V along with the mapping $V \times V \to \mathbb{C}$, $(u, v) \mapsto u \cdot v$, which satisfies for all vectors $u, v, w \in V$ and scalars $a \in \mathbb{C}$

(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 $u \cdot u > 0$ (notably if the expression is real).

Such mapping is called *scalar product* over V.

Real number $\sqrt{v \cdot v}$ is called *size of the vector* v and vector is *normalised*, if its size equals one. Vectors u and v are called *orthogonal* if their scalar product is zero, basis composed of mutually orthogonal and normalised vectors is called *orthonormal basis* V.

On first sight this is an extension of the definition of Euclidean vector spaces into the complex domain. We will keep on using the alternative notation $\langle u, v\rangle$ for scalar product of vectors u and v. Identically to the real domain, we obtain immediately from the definition the following simple properties of the scalar product for all vectors in V and scalars in $\mathbb{C}$:

$$\begin{gathered} u \cdot u \in \mathbb{R} \\ u \cdot u = 0 \quad \text{if and only if} \quad u = 0 \\ u \cdot (av) = \bar{a}(u \cdot v) \\ u \cdot (v + w) = u \cdot v + u \cdot w \\ u \cdot 0 = 0 \cdot u = 0 \\ \Big(\sum_i a_i u_i\Big) \cdot \Big(\sum_j b_j v_j\Big) = \sum_{i,j} a_i \bar{b}_j (u_i \cdot v_j), \end{gathered}$$

where the last equality holds for all finite linear combinations. It is a simple exercise to prove everything formally, for instance the fist property follows from the definition property (1).

Standard example of scalar product over complex vector space $\mathbb{C}^n$ is

$$(x_1, \dots, x_n)^T \cdot (y_1, \dots, x_n)^T = x_1 \bar{y}_1 + \dots + x_n \bar{y}_n.$$

Thanks to conjugation of the coordinates of the second argument this mapping satisfies all required properties. The space $\mathbb{C}^n$ with this scalar product is called *standard unitary space* of dimension n. We can denote this scalar product with matrix notation as $x \cdot y = \bar{y}^T \cdot x$.

the matrices A and C are primitive, and because

$$\begin{pmatrix}1/2 & 0 & 1/3\\ 0 & 1 & 1/2\\ 1/2 & 0 & 1/6\end{pmatrix}\cdot\begin{pmatrix}0\\1\\0\end{pmatrix}=\begin{pmatrix}0\\1\\0\end{pmatrix},$$

the middle column of the matrix B^n is always (for $n \in \mathbb{N}$) the vector $(0, 1, 0)^T$, that is, the matrix B cannot be primitive. The product

$$\begin{pmatrix}1/3 & 1/2 & 0 & 0\\ 1/2 & 1/3 & 0 & 0\\ 0 & 1/6 & 1/6 & 1/3\\ 1/6 & 0 & 5/6 & 2/3\end{pmatrix}\cdot\begin{pmatrix}0\\0\\a\\b\end{pmatrix}=\begin{pmatrix}0\\0\\a/6+b/3\\5a/6+2b/3\end{pmatrix},\quad a, b \in \mathbb{R}$$

implies that the matrix D^2 has in the right upper corner a zero two-dimensional (square) sub-matrix. By repetition of this implication we obtain that the same property is shared by the matrices $D^3 = D \cdot D^2$, $D^4 = D \cdot D^3, \dots, D^n = D \cdot D^{n-1}, \dots$, 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 realise that the powers of the permutation matrix are again permutation matrices. The matrix E is thus also not primitive. This can be easily verified by calculating the powers E^2, E^3, E^4. The matrix E^4 is a unit matrix. □

Now we show a more robust model.

3.22. Model of spreading of annual plants. We consider the plants that at the beginning of the summer blossom, at the peak of the summer produce seeds and die. 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 bigger in the spring and usually produce more seeds. After this, the whole cycle repeats.

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 the spring	small and big seedlings
beginning of the summer	small, medium and big blossoming flowers
peak of the summer	seeds
autumn	seedlings and seeds

We denote by $x_1(t)$ and by $x_2(t)$ the number of small and big seedlings respectively at the start of the spring in the year t and by $y_1(t)$, $y_2(t)$ and $y_3(t)$ the number of small, medium and big flowers respectively in the summer of that year. From the small seedlings either small or big flowers grow, from the big 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

Completely analogously to the Euclidean spaces and orthogonal mappings, great importance is in those mappings that respect scalar product.

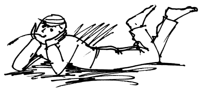
Unitary mapping

Linear mapping $\varphi : V \to W$ between unitary spaces is called *unitary mapping*, if for all vectors $u, v \in V$ we have

$$u \cdot v = \varphi(u) \cdot \varphi(v).$$

Unitary isomorphism is a bijective unitary mapping.

3.24. Properties of spaces with scalar product. In a brief discussion about Euclidean spaces in the previous chapter we have already derived some simple properties of spaces with scalar product. The proofs for the complex case are very similar.



In the following we shall work with real and complex spaces simultaneously and we write $\mathbb{K}$ for $\mathbb{R}$ or $\mathbb{C}$, in the real case the conjugation is just the identity mapping (as the actual restriction of the conjugation in the complex plane to the real line is). Similarly to the real space we define in general for arbitrary vector subspace $U \subset V$ in the space with scalar product its *orthogonal complement*

$$U^{\perp} = \{v \in V;\ u \cdot v = 0 \text{ for all } u \in U\},$$

which is clearly also a vector subspace in V.

In the following paragraphs we work exclusively with finitely-dimensional unitary or Euclidean spaces. However, many of our results have a natural generalisation for the so-called Hilbert spaces, which are specific infinitely-dimensional spaces with scalar products, to which we return later, albeit briefly.

Proposition. *For every finitely-dimensional space V of dimension n with scalar product we have:*

(1) *In V there exists an orthonormal basis.*
(2) *Every system of non-zero orthogonal vectors in V is linearly independent and can be extended to an orthogonal basis.*
(3) *For every system of linearly independent vectors $(u_1, \dots, u_k)$ there exists an orthonormal basis $(v_1, \dots, v_n)$ such that its vectors respectively generate the same subspaces as the vector u_j, that is, $\langle v_1, \dots, v_i\rangle = \langle u_1 \dots, u_i\rangle$, $1 \le i \le k$.*
(4) *If $(u_1, \dots, u_n)$ is an orthonormal basis V, then coordinates of every vector $u \in V$ are expressed via*

$$u = (u \cdot u_1)u_1 + \dots + (u \cdot u_n)u_n.$$

(5) *In any orthonormal basis the scalar product has the coordinate form*

$$u \cdot v = x \cdot y = x_1\bar{y}_1 + \dots + x_n\bar{y}_n$$

where x and y are columns of coordinates of the vectors u and v in a chosen basis. Notably, every n-dimensional space with scalar product is isomorphic to the standard Euclidean $\mathbb{R}^n$ or the unitary $\mathbb{C}^n$.
(6) *Orthogonal sum of unitary subspaces $V_1 + \dots + V_k$ in V is always a direct sum.*
(7) *If $A \subset V$ is an arbitrary subset, then $A^{\perp} \subset V$ is a vector (and thus also unitary) subspace and $(A^{\perp})^{\perp} \subset V$ is exactly the subspace generated by A. Furthermore we have $V = \langle A\rangle \oplus A^{\perp}$.*
(8) *V is orthogonal product of n one-dimensional unitary subspaces.*

i-th size, $i = 1, 2, 3$. Then we have

$$0 < b_{11} < 1, \quad b_{12} = 0, \quad 0 < b_{21} < 1, \quad 0 < b_{22} < 0, \quad b_{31} = 0,$$
$$0 < b_{32} < 1, b_{11} + b_{21} < 1, \quad b_{22} + b_{32} < 1$$

(think in detail about what each of these inequalities expresses). If we consider the classical probability, we can compute b_{11} as a ratio of the positive results (small seedling grew into a small flower) and of all possible results (the number of small seedlings), that is, $b_{11} = y_1(t)/x_1(t)$. From the

$$y_1(t) = b_{11}x_1(t).$$

Analogously, we obtain the equality

$$y_3(t) = b_{32}x_2(t).$$

If we denote for a while by $y_{2,1}(t)$ and $y_{2,2}(t)$ the number of medium flowers that grew out of small and big seedlings respectively, we have $y_2(t) = y_{2,1}(t) + y_{2,2}(t)$ and $b_{21} = y_{2,1}(t)/x_1(t)$, $b_{22} = y_{2,2}(t)/x_2(t)$ and thus

$$y_2(t) = b_{21}x_1(t) + b_{22}x_2(t).$$

Denote

$$B = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \\ 0 & b_{32} \end{pmatrix}, \qquad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \qquad y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix}$$

and rewrite the previous equation in the matrix notation

$$y(t) = Bx(t).$$

Denote by c_{11}, c_{12} and c_{13} the number of seed produced by small, medium and big flowers respectively, and by $z(t)$ the total number of produced seeds in the summer of the year t, we have

$$z(t) = c_{11}y_1(t) + c_{12}y_2(t) + c_{13}y_3(t),$$

or 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.

Denote finally 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, and by d_{11} and d_{21} the probabilities that the seed burst out in the autumn and that the seed does not burst respectively, and by f_{11} and f_{22} the probabilities that the seedling and the seed do

Proof. (1), (2), (3): We first extend the given system of vectors into any basis $(u_1, \dots, u_n)$ of the space V and then start the Gramm-Schmidt orthogonalisation from 2.42. This yields an orthogonal basis with properties as required in (3). But from the Gramm-Schmidt orthogonalisation algorithm it is clear that when the original k vectors formed an orthogonal system of vectors, then during the process they won't be changed. Thus we have also proved (2) and (1).
(4): If $u = a_1u_1 + \dots + a_nu_n$, then

$$u \cdot u_i = a_1(u_1 \cdot u_i) + \dots + a_n(u_n \cdot u_i) = a_i\|u_i\|^2 = a_i$$

(5): Similarly we compute for any vectors $u = x_1u_1 + \dots + x_nu_n$, $v = y_1u_1 + \dots + y_nu_n$

$$\begin{aligned} u \cdot v &= (x_1u_1 + \dots + x_nu_n) \cdot (y_1u_1 + \dots + y_nu_n) \\ &= x_1\bar{y}_1 + \dots + x_n\bar{y}_n. \end{aligned}$$

(6): We need to show that for any tuple V_i, V_j from the given subspaces their intersection is trivial. If we have $u \in V_i$ and simultaneously $u \in V_j$, then we have $u \perp u$, that is, $u \cdot u = 0$. That 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). We have thus checked that $A^\perp$ is a unitary subspace in V. Let $(v_1, \dots, v_k)$ be some basis $\langle A \rangle$ chosen among the elements of A, and let be $(u_1, \dots, u_k)$ the orthonormal basis outputted by the Gramm-Schmidt orthogonalisation of the vectors $(v_1, \dots, v_k)$. We extend it to the orthonormal basis of the whole V (both exist thanks to the already proven parts of this proposition). Because it is an orthogonal basis, necessarily $\langle u_{k+1}, \dots, u_n \rangle = \langle u_1, \dots, u_k \rangle^\perp = A^\perp$ and $A \subset \langle u_{k+1}, \dots, u_n \rangle^\perp$ (this follows from expressing the coordinates under the orthonormal bassi). If $u \perp \langle u_{k+1}, \dots, u_n \rangle$, then u is necessarily a linear combination of the vectors $u_1, \dots, u_k$, but that happens whenever it is a linear combination of the vectors $v_1, \dots, v_k$, which is equivalent to u being in $\langle A \rangle$.
(8): This is equivalent to the formulation of existence of the orthonormal basis. □

3.25. Important properties of size. Now we have everything prepared for basic properties connected with our definition of the size of vectors. We speak also of the *norm* defined by the scalar product. Note also that all claims always consider finite sets of vectors and their validity does not depend on the dimension of the space V where it all takes place.

Theorem. *For any two vectors u, v in the space V with scalar product we have*

1. *$\|u + v\| \le \|u\| + \|v\|$, with equality if and only if u and v are linearly dependent.*
   (triangle inequality)
2. *$|u \cdot v| \le \|u\|\,\|v\|$, with equality if and only if u and v are linearly dependent.*
   (Cauchy inequality)
3. *For every orthonormal system of vectors $(e_1, \dots, e_k)$ we have*

$$\|u\|^2 \ge |u \cdot e_1|^2 + \dots + |u \cdot e_k|^2$$

(Bessel inequality).

not die during the winter respectively. The probabilities d_{11}, d_{21} clearly must satisfy the inequalities

$$0 < d_{11}, \quad 0 < d_{21}, \quad d_{11} + d_{21} = 1,$$

and because a seedling dies in the winter more easily than a seed hidden in the ground, we assume about f_{11}, f_{22} 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 for the numbers in individual stages of flowers in the following year from the previous equalities compose the recurrent formulas:

$$\begin{aligned} x(t+1) =& Fw(t) = F\big(Dz(t)\big) = (FD)z(t) = (FD)\big(Cy(t)\big) = \\ =& (FDC)y(t) = (FDC)\big(Bx(t)\big) = (FDCB)x(t), \end{aligned}$$

$$\begin{aligned} y(t+1) =& Bx(t+1) = B\big(Fw(t)\big) = (BF)w(t) = (BF)\big(Dz(t)\big) = \\ =& (BFD)z(t) = (BFD)\big(Cy(t)\big) = (BFDC)y(t), \end{aligned}$$

$$\begin{aligned} z(t+1) =& Cy(t+1) = C\big(Bx(t+1)\big) = (CB)x(t+1) = (CB)\big(Fw(t)\big) = \\ =& (CBF)w(t) = (CBF)\big(Dz(t)\big) = (CBFD)z(t), \end{aligned}$$

$$\begin{aligned} w(t+1) =& Dz(t+1) = D\big(Cy(t+1)\big) = (DC)y(t+1) = \\ =& (DC)\big(Bx(t+1)\big) = (DCB)x(t+1) = (DCB)\big(Fw(t)\big) = \\ =& = (DCBF)w(t). \end{aligned}$$

Using the notation

$$A_x = FDCB, \quad A_y = BFDC, \quad A_z = CBFD, \quad A_w = DCBF,$$

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

$$\begin{aligned} x(t) =& A_x x(t-1) = A_x^2 x(t-2) = \cdots = A_x^{t-1} x(1) = A_x^{t-1} Fw(0) = \\ =& A_x^{t-1} FDz(0). \end{aligned}$$

(4) *For orthonormal system of vectors $(e_1, \dots, e_k)$ the vector u belongs to the subspace $\in \langle e_1, \dots, e_k \rangle$ if and only if*

$$\|u\|^2 = |u \cdot e_1|^2 + \cdots + |u \cdot e_k|^2.$$

(Parseval equality)

(5) *For orthonormal system of vectors $(e_1, \dots, e_k)$ and vector $u \in V$ is the vector*

$$w = (u \cdot e_1)e_1 + \cdots + (u \cdot e_k)e_k$$

the only vector which minimises the size $\|u - v\|$ for all $v \in \langle e_1, \dots, e_k \rangle$.

Proof. All proofs rely on direct computations:
(2): Define the vector $w := u - \frac{u \cdot v}{v \cdot v} v$, that is, $w \perp v$ and compute

$$0 \le \|w\|^2 = \|u\|^2 - \tfrac{(\overline{u \cdot v})}{\|v\|^2}(u \cdot v) - \tfrac{u \cdot v}{\|v\|^2}(v \cdot u) + \tfrac{(u \cdot v)(\overline{u \cdot v})}{\|v\|^4}\|v\|^2$$

$$0 \le \|w\|^2\|v\|^2 = \|u\|^2\|v\|^2 - 2(u \cdot v)(\overline{u \cdot v}) + (u \cdot v)(\overline{u \cdot v})$$

>From there it directly follows that $\|u\|^2\|v\|^2 \ge |u \cdot v|^2$ and the equality holds if and only if $w = 0$, that is, whenever u and v are linearly dependent.
(1): Again it suffices to compute

$$\begin{aligned} \|u + v\|^2 &= \|u\|^2 + \|v\|^2 + u \cdot v + v \cdot u \\ &= \|u\|^2 + \|v\|^2 + 2\,\mathrm{Re}(u \cdot v) \\ &\le \|u\|^2 + \|v\|^2 + 2|u \cdot v| \le \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \\ &= (\|u\| + \|v\|)^2 \end{aligned}$$

Because these are positive real numbers, it indeed is that $\|u+v\| \le \|u\| + \|v\|$. Furthermore, with equality it must be that in all previous inequalities equality also holds, but that is equivalent to the condition that u and v are linearly dependent (using the previous part).
(3), (4): Let $(e_1, \dots, e_k)$ be an orthonormal system of vectors. We extend to an orthonormal basis $(e_1, \dots, e_n)$ (that is always possible thanks to the previous theorem). Then, again using the previous theorem, is for every vector $u \in V$

$$\|u\|^2 = \sum_{i=1}^{n} (u \cdot e_i)(\overline{u \cdot e_i}) = \sum_{i=1}^{n} |u \cdot e_i|^2 \ge \sum_{i=1}^{k} |u \cdot e_i|^2$$

But that is the Bessel inequality. Furthermore, equality can hold if and only if $u \cdot e_i = 0$ for all $i > k$, which proves the Parseval equality.
(5): Choose arbitrary $v \in \langle e_1, \dots, e_k \rangle$ and extend the given orthonormal system to the orthonormal basis $(e_1, \dots, e_n)$. Let $(u_1, \dots, u_n)$ and $(x_1, \dots, x_k, 0, \dots, 0)$ be coordinates of u and v under this basis. Then

$$\|u - v\|^2 = |u_1 - x_1|^2 + \cdots + |u_k - x_k|^2 + |u_{k+1}|^2 + \cdots + |u_n|^2$$

and this expression is clearly minimised when choosing the individual vectors to be $x_1 = u_1, \dots, x_k = u_k$. □

3.26. Properties of unitary spaces. The properties of orthogonal mapping have a direct analogue in the complex domain. We can easily formulate them and prove together:

Proposition. *Consider the linear mapping (endomorphism) $\varphi : V \to V$ on the space with scalar product. Then the following conditions are equivalent.*

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 $\lambda = A_z$, compute

(3.5)

$$\lambda = CBFD = \begin{pmatrix} c_{11} & c_{12} & c_{13}\end{pmatrix}\begin{pmatrix} b_{11} & 0\\ b_{21} & b_{22}\\ 0 & b_{32}\end{pmatrix}\begin{pmatrix} f_{11} & 0\\ 0 & f_{22}\end{pmatrix}\begin{pmatrix} d_{11}\\ d_{21}\end{pmatrix} =$$

$$= \begin{pmatrix} c_{11}b_{11} + c_{12}b_{21} & c_{12}b_{22} + c_{13}b_{32}\end{pmatrix}\begin{pmatrix} f_{11}d_{11}\\ f_{22}d_{21}\end{pmatrix} =$$

$$= b_{11}c_{11}d_{11}f_{11} + b_{21}c_{12}d_{11}f_{11} + b_{22}c_{12}d_{21}f_{22} + b_{32}c_{13}d_{21}f_{22}$$

and order the previous computation into a suitable form

$$\begin{aligned} x(t) &= (FDCB)^{t-1}FDz(0) = FD(CBFD)^{t-2}CBFDz(0) =\\ &= FD(CBFD)^{t-1}z(0) = FDA_z^{t-1}z(0) = \lambda^{t-1}FDz(0); \end{aligned}$$

in this way only two matrix multiplications remain.

Let us 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 $\lambda = A_z = 1.3325$ expresses the relative increment of the population between two years. Check by yourself that each of the matrices A_x, A_y, A_w has only one non-zero eigenvalue $\lambda = 1.3325$; other eigenvalues are equal to 0.

We show one more application of the given model. We can be 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. Let us be more precise in the formulation of the question. By *flexibility of the reaction of the characteristic λ on the parameter s*, denote d 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

(1) φ is unitary or orthogonal transformation,
(2) φ is linear isomorphism and for every $u, v \in V$ it holds that $\varphi(u)\cdot v = u\cdot\varphi^{-1}(v)$,
(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) matrix A of a mapping φ under some orthonormal basis satisfies $A^{-1} = \bar{A}^T$,
(5) rows of the matrix A of the mapping φ under orthonormal basis form an orthonormal basis of the space $\mathbb{K}^n$ with standard scalar product,
(6) columns of the matrix A of the mapping φ under orthonormal basis form an orthonormal basis of the space $\mathbb{K}^n$ with standard scalar product.

Proof. (1) ⇒ (2): The mapping φ is injective, therefore it must be onto. It also holds that $\varphi(u)\cdot v = \varphi(u)\cdot\varphi(\varphi^{-1}(v)) = u\cdot\varphi^{-1}(v)$.
(2) ⇒ (3): Standard scalar product is in $\mathbb{K}^n$ always given for columns x, y of scalars by the expression $x\cdot y = x^T E\bar{y}$, where E is the unit matrix. The property (2) thus means that the matrix A of the mapping φ is invertible and it holds that $(Ax)^T\bar{y} = x^T\overline{A^{-1}y}$. That means that $\bar{x}^T(\bar{A}^Ty - A^{-1}y) = 0$ for all $x \in \mathbb{K}^n$. Notably by substituting the expression in the parentheses for x we find out that this is possible only when $\bar{A}^T = A^{-1}$.
(3) ⇔ (4): If $\bar{A}^T = A^{-1}$ under some orthonormal basis, then the condition (2) holds ($\varphi(u)\cdot v = (Ax)^T E\bar{y} = x^T E\overline{A^{-1}y} = u\cdot\varphi^{-1}(v)$) and thus also (3).
(4) ⇒ (5) The claim is expressed via the matrix A of the mapping φ as the equation $A\bar{A}^T = E$, which is ensured thanks to (4).
(5) ⇒ (6): Because for the determinant we have $|\bar{A}^TA| = |E| = |A\bar{A}^T| = |A||\overline{A}| = 1$, there exists the inverse matrix A^{-1}. But we also have $A\bar{A}^TA = A$, therefore also $\bar{A}^TA = E$ which is expressed exactly by (6).
(6) ⇒ (1): In the chosen orthonormal basis we have

$$\varphi(u)\cdot\varphi(v) = (Ax)^T\overline{(Ay)} = xA^T\bar{A}\bar{y} = x^T\bar{E}\bar{y} = x^T\bar{y}$$

where x and y are columns of coordinates of the vectors u and v. That ensures that the scalar product is preserved. □

Characterisation from the previous theorem deserves some notes. The matrix $A \in \mathrm{Mat}_n(\mathbb{K})$ with the property $A^{-1} = \bar{A}^T$ are called *unitary matrices* for complex scalars (and in the case $\mathbb{R}$ we have already used the name *orthogonal matrices* for them). From the definition we have that a product of unitary (orthogonal) matrices is again unitary (orthogonal), the same holds 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 *unitary group* and of *orthogonal group*.

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 always size equal to one, in the case of real scalars the determinant is equal ±1. Furthermore, if $Ax = \lambda x$ for 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 equal ±1, the

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 (||3.5||)

$$\lambda(f_{22}) = d_{21}(b_{22}c_{12}+b_{32}c_{13})f_{22}+d_{11}(b_{11}c_{11}f_{11}+b_{21}c_{12}f_{11})$$

and for specific values of the other parameters

$$\lambda(f_{22}) = 13f_{22}+0.0325.$$

Because $f_{22} = 0.1$, we can compute

$$\lambda(0.1) = 1.3325, \quad \lambda(0.1+\Delta s) = 1.3325+13\Delta s, \quad \Delta\lambda(0.1) = 13\Delta s,$$

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 seed that overwinter (parameter d_{21}) and their survivability (parameter f_{22}). This revelation is not surprising, the farmers are aware of this fact since the times of neolithic times. The result shows that the mathematical model indeed adequately describes the reality.

Other interesting and well-described models of growth can be found in the collection of exercises after this chapter.

3.23. Consider the following Leslie model: farmer breeds sheep. The birth-rate of sheep depends only on their age and is on average 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. Farmer would like to sell (living) lambs younger than one year for their skin. What part of the lambs can be sold every year such that the

eigenvalues of unitary matrices are always complex units in the complex plane.

As with the orthogonal mappings we can easily check that orthogonal complements of invariant subspaces with respect to unitary $\varphi : V \to V$ are always 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 must thus be a bijection, notably we have that $\varphi^{-1}(u) \in U$. But then $\varphi(v)\cdot u = 0$, because $v \in U^\perp$. That means that also $\varphi(v) \in U^\perp$.

This yields 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 orthogonal sum of one-dimensional eigensubspaces.*

Proof. There surely exist at least one eigenvector $v \in V$. Then the restriction φ on the invariant subspace $\langle v\rangle^\perp$ is again unitary and surely has also some 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 easily understand the details of the proof of spectral decomposition of the orthogonal mapping from 2.50 at the end of the second chapter – real matrix of an orthogonal mapping is interpreted as a matrix of a unitary mapping on a complex extension of Euclidean space and we carefully observe the corollaries of the structure of the roots of the real characteristic polynomial over the complex domain. We automatically obtain invariant two-dimensional subspaces given by tuples of complexly conjugated eigenvalues and thus the corresponding rotation for restricted original real mapping.

3.27. Dual and adjoint mappings. When discussing vector spaces and linear mappings in the second chapter we have already briefly mentioned the dual vector space V^* of all linear forms over the vector space V, see 2.39.

For every linear mapping between vector spaces $\psi : V \to W$ we can naturally define its *dual mapping* $\psi^* : W^* \to V^*$ by the relation

$$\langle v, \psi^*(\alpha)\rangle = \langle\psi(v), \alpha\rangle, \tag{3.6}$$

where $\langle\ ,\ \rangle$ denotes evaluation of the form (the second argument) on the vector (first argument), $v \in V$ and $\alpha \in W^*$ are arbitrary.

Let us choose bases $\underline{v}$ over V, $\underline{w}$ over W and let us write A for the matrix of the mapping ψ under these bases. Then we easily compute in dual bases the matrix of the mapping ψ^* in the corresponding dual bases over the dual spaces. Indeed, the definition says that if we represented 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:

$$\langle\psi(v), \alpha\rangle = (\alpha_1, \dots, \alpha_n)\cdot A\cdot\begin{pmatrix} v_1\\ \vdots\\ v_n\end{pmatrix} = \langle v, \psi^*(\alpha)\rangle.$$

That means that the matrix of the dual mapping ψ^* is the transpose A^T, because $\alpha\cdot A = (A^T\cdot\alpha^T)^T$.

size of the herd remains the same? In what ratio will then the sheep be distributed among individual age categories?

Solution. 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

$$L = \begin{pmatrix} 0 & 2 & 5 & 2 \\ a & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix},$$

and 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

$$\lambda^4 - 2a\lambda^2 - \frac{5}{2}\lambda - \frac{1}{2},$$

and if we require it to have 1 as a root, it must be that $a = \frac{1}{5}$ (we substitute $\lambda = 1$ and set the polynomial equal to zero). 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. □

3.24. 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). ○

D. Markov processes

3.25. Sweet-toothed gambler. 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 sweet and if he wins, he gains one additional, 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?

Let us further assume that we are in a vector space with scalar product. If we choose one fixed vector $v \in V$, substituting vectors for the second argument in the scalar product gives us a mapping $V \to V^* = \mathrm{Hom}(V, \mathbb{K})$

$$V \ni v \mapsto (w \mapsto \langle v, w \rangle \in \mathbb{K}).$$

The non-degeneracy condition of the scalar product ensures that this mapping is a bijection. Furthermore we know that it indeed is a linear mapping over complex or real scalars, because we have fixed the second argument. On first sight it is clear that the vectors of the orthonormal basis are mapped on forms that constitute a dual basis, and every vector can be thus understood using the scalar product as a linear form.

In the case of vector spaces with scalar product our identification of a vector space with its dual also takes the dual mapping ψ^* to the mapping $\psi^* : W \to V$ given by the formula

$$\langle \psi(u), v \rangle = \langle u, \psi^*(v) \rangle, \tag{3.7}$$

where by the same notation of parentheses as in the definition (3.6) we now mean scalar product. This mapping is called *adjoint mapping* to ψ.

Equivalently we can understand the relation (3.27) to be the definition of the adjoint mapping ψ^*, for instance by substituting all tuples of vectors of an orthonormal basis for the vectors u and v we directly obtain all values of the matrix of the mapping ψ^*.

The previous calculation for the dual mapping in coordinates can be now repeated, we just have to keep in mind that in orthonormal bases in unitary spaces the coordinates of the second argument are conjugated:

$$\begin{aligned} \langle \psi(v), w \rangle &= \overline{(w_1, \dots, w_n)} \cdot A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\ &= \overline{\left(\bar{A}^T \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \right)^T} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \langle v, \psi^*(w) \rangle \end{aligned}$$

Therefore we see 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 adjoint matrices are well defined for any rectangular matrix. We should not confuse them with algebraic adjoints, which we have used for square matrices when working with determinants.

We can thus summarise that for any linear mapping $\psi : V \to W$ between unitary spaces under orthonormal bases with the matrix A, its dual mapping has in the dual bases the matrix A^T. If we also identify using the scalar product the vector spaces with their duals, then the dual mapping corresponds to the adjoint mapping $\psi^* : W \to V$ (it is a custom to denote this mapping in the same way as the dual mapping), which has the matrix A^*. The distinction between the matrix of the dual mapping and of the adjoint mapping is thus in the additional conjugation, which is of course the corollary of the fact that unifying the unitary space with its dual is not complexly linear mapping (since from the second argument in the scalar product the scalars are brought out conjugated).

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 with $1/2$ probability $(i-1)$ sweets with $1/2$ probability $(i+1)$ sweets. 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 we have

$$X_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

after four bets the situation is described by the following vector

$$X_5 = A^4 X_1 = \begin{pmatrix} \frac{1}{8} \\ \frac{3}{16} \\ 0 \\ \frac{5}{16} \\ 0 \\ \frac{3}{8} \end{pmatrix},$$

that is, the probability that the game ends in the fourth bet or sooner is one half.

Note also 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 and 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 – the case that the player has no sweet, that is $x = (1, 0, 0, 0, 0, 0)^T$, or 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

3.28. Self-adjoint mappings. A special case of linear mapping are those that are identical with their adjoint mappings: $\psi^* = \psi$. Such mappings are called *self-adjoint*. Equivalently we can say that they are those mappings whose matrix A is under one (and thus under all) orthogonal basis satisfies $A = A^*$.

In the case of Euclidean spaces the self-adjoint mappings are those that have symmetric matrix (under some basis). Often they are called *symmetric matrices* and *symmetric mappings*.

In the complex domain the matrices that satisfy $A = A^*$ are called *Hermitian 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 subspace in the complex domain.

Remark. Especially interesting is in this connection the following remark. 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. As every real matrix is a sum of a symmetric and an anti-symmetric part,

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T),$$

in the complex domain we analogously have

$$A = \frac{1}{2}(A + A^*) + i\frac{1}{2i}(A - A^*)$$

and can thus express every complex matrix in exactly one way as a sum

$$A = B + i\,C$$

with Hermitian matrices B and C. It 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 using two self-adjoint mappings.

3.29. Spectral decomposition. We consider a self-adjoint mapping $\psi : V \to V$ with matrix A under some orthonormal basis and we try to proceed similarly as in 2.50. Again, we first look in general at the invariant subspaces of self-adjoint mappings and on their orthogonal complements. If for any subspace $W \subset V$ and self-adjoint mapping $\psi : V \to V$ we have $\psi(W) \subset W$, then also for every $v \in W^\perp$, $w \in W$

$$\langle \psi(v), w \rangle = \langle v, \psi(w) \rangle = 0.$$

That means that also $\psi(W^\perp) \subset W^\perp$.

Consider now the matrix A of a self-adjoint mapping under some orthonormal basis and $A \cdot x = \lambda x$ for some eigenvector $x \in \mathbb{C}^n$. We obtain

$$\lambda\langle x, x \rangle = \langle Ax, x \rangle = \langle x, Ax \rangle = \langle x, \lambda x \rangle = \bar{\lambda}\langle x, x \rangle.$$

Positive real number $\langle x, x \rangle$ can be cancelled out and thus it must be $\bar{\lambda} = \lambda$, that is, eigenvalues are always real.

The characteristic polynomial $\det(A - \lambda E)$ has that many complex roots as is the dimension of the square matrix A, and all of them are actually real. Thus we have proved important general result:

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. □

3.26. Car rental. 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 some long time, if the cars are distributed at the start in a random way?

Solution. Let us 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 procentual distribution of the cars. At the end of the week is according to the statement the state 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 vector x for which it holds 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 $(0.1 - \lambda)(0.8 - \lambda) - 0.9.0.2 = (\lambda - 1)(\lambda + 0.1)$ 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 procentual 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 week is described by the vector $x_n = A^n x$. Now it is useful to express the initial vector x in the basis of the eigenvectors of A. The eigenvector of the eigenvalue 1 has already been found and similarly we find eigenvectors of the eigenvalue -0.1. That is for instance the vector $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

Proposition. *Orthogonal complement of an invariant subspace for self-adjoint mapping is also invariant. Furthermore, all eigenvalues of a Hermitian matrix A are always real.*

>From the definition itself it is clear that restriction of a self-adjoint mapping on an invariant subspace is again self-adjoint. The previous claim thus ensures that there always exists a basis of V composed of eigenvectors. Indeed, the restriction of ψ on the orthogonal complement of an invariant subspace is again self-adjoint mapping, thus we can add into the basis one eigenvector after another, until we obtain whole decomposition of V. Eigenvector associated with distinct eigenvalues are perpendicular, because from the equations $\psi(u) = \lambda u$, $\psi(v) = \mu v$ we have that

$$\lambda\langle u, v\rangle = \langle \psi(u), v\rangle = \langle u, \psi(v)\rangle = \bar{\mu}\langle u, v\rangle = \mu\langle u, v\rangle.$$

Usually our result is formulated using projections on eigensubspaces. About the projector $P : V \to V$ we say that it is *perpendicular* if $\operatorname{Im} P \perp \operatorname{Ker} P$. Two perpendicular projectors P, Q are *mutually perpendicular* if $\operatorname{Im} P \perp \operatorname{Im} Q$.

Theorem (About the 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 $\lambda_1, \dots, \lambda_k$ are all distinct eigenvalues of ψ and $P_1, \dots, P_k$ are the corresponding perpendicular and mutually perpendicular projectors on the eigenspaces corresponding to the eigenvalues, then*

$$\psi = \lambda_1 P_1 + \dots + \lambda_k P_k.$$

Dimension of images of these projectors is always equal to the algebraic multiplicity of the eigenvalues λ_i.

3.30. Orthogonal diagonalisation. Mappings for which we can find an orthonormal basis as in the previous theorem about spectral decomposition are called *orthogonally diagonalisable*. They are of course exactly such mappings for which we can find an orthonormal basis such that the matrix of the mapping is diagonal under this basis. Let us think for a while how can they look like.

For the Euclidean case it is simple: diagonal matrices are first of all symmetric, thus they are exactly the self-adjoint mappings. As a corollary we obtain a result that an orthogonal mapping of an Euclidean space into itself is orthogonally diagonalisable if and only if it is self-adjoint (they are exactly the self-adjoint mappings with eigenvalues ±1).

For complex unitary spaces the situation is more complicated. Consider arbitrary linear mapping $\varphi : V \to V$ of a unitary space and let $\varphi = \psi + i\eta$ be the (uniquely given) decomposition of φ into its Hermitian and anti-Hermitian part. If φ has under a suitable orthonormal basis a diagonal matrix D, then $D = \operatorname{re} D + i \operatorname{im} D$, where the real and the imaginary parts are exactly the matrices ψ and η (follows from the uniqueness of the decomposition). Thus it also holds that $\psi \circ \eta = \eta \circ \psi$ and $\varphi \circ \varphi^* = \varphi^* \circ \varphi$. The mappings $\varphi : V \to V$ with the last listed property are called *normal*.

Mutual connections are shown in the following proposition (we follow the notation of this paragraph):

Proposition. *The following conditions are equivalent:*

(1) φ is orthogonally diagonalisable,
(2) $\varphi^* \circ \varphi = \varphi \circ \varphi^*$ (that is, φ is a normal mapping),
(3) $\psi \circ \eta = \eta \circ \psi$,

The initial vector can thus be expressed as a linear combination $x = a\begin{pmatrix}0.2\\0.9\end{pmatrix} + b\begin{pmatrix}-1\\1\end{pmatrix}$. 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 for $n \to \infty$ approaching zero and 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 *and* can be easily expressed using the initial states of the cars: $a = \frac{x_B+x_P}{1,1}$. □

3.27. **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, find its eigenvalues and eigenvectors. ○

3.28. Students at the lecture. Students can be divided into, say, three group – those that are present on a lecture and pay attention, those that are present but pay no attention and those who are in a pub instead. Now let us 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. Let us say that it can be as follows:

Student that pays attention: with probability 50% stays in the same state, with 40% stops paying attention and with 10% moves to the pub. Student that pays no attention: starts paying attention with 10%, with 50% stays in the same state and with 40% moves to the pub. Student that is in 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 $\begin{pmatrix}-0.5 & 0.1 & 0\\0.4 & -0.5 & 0\\0.1 & 0.4 & 0\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix} = 0$, which are exactly multiples of the vector $(0, 0, 1)$. In other words, all students end up in the pub.

(4) *for a matrix $A = (a_{ij})$ of a mapping φ under some orthonormal basis and its $m = \dim V$ eigenvalues λ_i it holds that $\sum_{i,j} |a_{ij}|^2 = \sum_{i=1}^m |\lambda_i|^2$.*

Brief proof. The implication (1) ⇒ (2) was already dealt with.

(2) ⇔ (3): it suffices to do a direct calculation

$$\varphi\varphi^* = (\psi + i\eta)(\psi - i\eta) = \psi^2 + \eta^2 + i(\eta\psi - \psi\eta)$$
$$\varphi^*\varphi = (\psi - i\eta)(\psi + i\eta) = \psi^2 + \eta^2 + i(\psi\eta - \eta\psi)$$

Subtraction yields $2i(\eta\psi - \psi\eta)$.

(2) ⇒ (1): let $u \in V$ be an eigenvector of the normal mapping φ. Then

$$\varphi(u)\cdot\varphi(u) = \langle\varphi^*\varphi(u), u\rangle = \langle\varphi\varphi^*(u), u\rangle = \varphi^*(u)\cdot\varphi^*(u),$$

thus also $|\varphi(u)| = |\varphi^*(u)|$. If φ is normal, then $(\varphi - \lambda\,\mathrm{id}\,V)^* = (\varphi^* - \bar\lambda\,\mathrm{id}\,V)$ and thus $(\varphi - \lambda\,\mathrm{id}\,V)$ is also a normal mapping. From the previous equation follows that if $\varphi(u) = \lambda u$, then $\varphi^*(u) = \bar\lambda u$. That means that φ and φ^* have the same eigenvectors and conjugated eigenvalues.

As with self-adjoint mappings we know easily prove orthogonal diagonalisability. A necessary and sufficient condition for that is that the orthogonal complement of every eigensubspace for normal φ is invariant (note that a restriction of a normal mapping on an invariant subspace is again normal). Consider an eigenvector $u \in V$ with eigenvalue λ, $v \in \langle u\rangle^\perp$. We have

$$\varphi(v)\cdot u = v\cdot\varphi^*(u) = \langle v, \bar\lambda u\rangle = \lambda u\cdot v = 0$$

and thus again $\varphi(v) \in \langle u\rangle^\perp$.

(1) ⇔ (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 it does not depend on the choice of orthonormal basis. Thus if φ is diagonalisable, this expression equals exactly $\sum_i|\lambda_i|^2$.

The other implication is a direct corollary of the Schur theorem about unitary triangulation of an arbitrary linear mapping $V \to V$, which we prove later in 3.37. The theorem says that for every linear mapping $\varphi : V \to V$ there exists an orthonormal basis under which φ has upper triangular matrix. On its diagonal there must be then all the eigenvalues of φ. As we have already shown, the expression $\sum_{i,j}|a_{ij}|^2$ does not depend on the choice of the orthonormal bassi, thus from the assumed equality we have that all elements that are not on the diagonal must be in this matrix equal to zero. □

In terms of matrices of mappings we obtain: a mapping is normal if and only if its matrix under some orthonormal basis (equivalently, under any orthonormal basis) satisfies $AA^* = A^*A$. Such matrices are called *normal matrices*.

Remark. Note that for the calculations with linear mappings on complex unitary spaces, we can understand the last theorem as a generalisation of common calculations with complex numbers in the polar form – the role of real numbers is played by self-adjoint mappings, the role of complex numbers is played by unitary mappings. Very notable is also the analogy to the expression of complex units in the form $\cos t + i\sin t$ with the property $\cos^2 t + \sin^2 t = 1$:

Corollary. *Unitary mappings on unitary space V are exactly those normal mappings for which the aforementioned unique decomposition $\varphi = \psi + i\eta$ satisfies $\psi^2 + \eta^2 = \mathrm{id}\,V$.*

Such 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 we have that the state stabilises on the eigenvector associated with the eigenvalue 1. That is in this case the solution of the equation

$$\begin{pmatrix}-0.5 & 0.1 & 0\\ 0.4 & -0.5 & 0.1\\ 0.1 & 0.4 & -0.1\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = 0.$$

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 for 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. □

3.29. Roulette. The player of the roulette has the following strategy: he came 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 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 easily construct the matrix:

$$A = \begin{pmatrix}1 & a & a & a & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & b & 0 & 0 & 0\\ 0 & 0 & b & 0 & 0\\ 0 & 0 & 0 & b & 1\end{pmatrix},$$

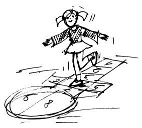
where $a = \frac{19}{37}$ and $b = \frac{18}{37}$. Note that the matrix is probabilistic and singular. The eigenvalue 1 is double. The game does not converge to a single vector x_∞, but ends in one of the eigenvectors associated with eigenvalue 1, that is, either $(1, 0, 0, 0, 0)$ (the player looses 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 $n \geq 3$:

$$A^\infty := A^3 = A^n = \begin{pmatrix}1 & a+ab+ab^2 & a+ab & a & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & b^3 & b^2 & b & 1\end{pmatrix}$$

and 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

Proof. For unitary mapping φ is $\varphi\varphi^* = \mathrm{id}\, V = \varphi^*\varphi$ and thus $\varphi\varphi^* = (\psi + i\eta)(\psi - i\eta) = \psi^2 + 0 + \eta^2 = \mathrm{id}\, V$. On the other hand, for normal mapping the last calculation shows that the other implication holds too. □

3.31. Non-negative mappings and roots. Non-negative real numbers are exactly those which we can write as square roots. Generalisation of such behaviour for matrices and mappings can be seen in products of matrices $B = A^* \cdot A$ (that is, composition of mappings $\psi^* \circ \psi$):



$$\langle B\cdot x, x\rangle = \langle A^*\cdot A\cdot x, x\rangle = \langle A\cdot x, A\cdot x\rangle \geq 0$$

for all vectors x. Furthermore, we clearly have

$$B^* = (A^*\cdot A)^* = A^*\cdot A = B.$$

Hermitian matrices B with such property are called *positively semidefinite* and if the zero value is attained only for $x = 0$, they are called *positively definite*. Analogously, we speak of *positively definite* and *positively semidefinite* mappings $\psi : V \to V$.

For every positively semidefinite mapping $\psi : V \to V$ we can find its root, that is, a mapping η such that $\eta\circ\eta = \psi$. It is simplest to see under an orthonormal basis where ψ has diagonal matrix. Such basis exists (as we have already proven) and the matrix A of the mapping ψ has on diagonal only non-negative real numbers, the eigenvalues of ψ. If some of them were negative, then the condition for non-negativity would not be satisfied already for some of the basis vectors. But then it suffices to define the mapping η using the matrix B with square roots of the corresponding eigenvalues on diagonal.

3.32. Spectra and nilpotent mappings. At the end of this section we return to the question about behaviour of linear mapping in full generality. We shall still work with real or complex vector spaces.

Let us recall that *spectrum of linear mapping* $f : V \to V$ is a sequence of roots of the characteristic polynomial of the mapping f, counting multiplicities. *Algebraic multiplicity* of eigenvalue is its multiplicity as of a root of the characteristic polynomial, *geometric multiplicity* of eigenvalue is the dimension of the corresponding subspace of eigenvectors.

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 property is called *degree of nilpotency* of the mapping f. The mapping $f : V \to V$ is called *cyclic*, if there exists a 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 under this basis is of the form

$$A = \begin{pmatrix}0 & 1 & 0 & \dots\\ 0 & 0 & 1 & \dots\\ \vdots & \vdots & & \ddots\end{pmatrix}.$$

If $f(v) = a\cdot v$, then for every natural k we have $f^k(v) = a^k\cdot v$. Notably, the spectrum of nilpotent mapping can contain only zero scalar (and that is always present).

Directly from the definition follows that every cyclic mapping is nilpotent, furthermore its degree of nilpotency is equal to the as a win of €80. (We multiply by the matrix A^∞ the initial vector $(0, 1, 0, 0, 0)$ and obtain the vector $(a + ab + ab^2, 0, 0, 0, b^3)$.) □

3.30. 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}. ○

3.31. Absent-minded professor. Consider the following situation: absent-minded professor carries an umbrella with him, but with probability $1/2$ he forgets it where he is leaving from. In the morning, he leaves to the work. At the work, he goes for a lunch into a restaurant, and then back. After he is finished with his work, he leaves for home. Consider for simplicity that he does not go anywhere else and that in the restaurant the umbrella stays on his favourite spot, where he can take it from on the next day (if he does not forget it there). Consider this situation as 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 useful to take as a time unit one day – from morning to morning.)

Solution.

$$A = \begin{pmatrix} 11/16 & 3/8 & 1/4 \\ 3/16 & 3/8 & 1/4 \\ 1/8 & 1/4 & 1/2 \end{pmatrix}$$

We compute for instance the element a_1^1, that is, the probability that the umbrella starts its day at home and stays there (that is, will be there the next day in the morning) – there are three disjoint ways for the umbrella:

D the professor forgets it at home in the morning $p_1 = \frac{1}{2}$,

DPD the professor takes it to the work, then he forgets to take it on the 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 all the time with him 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 $1/(2 + 1 + 1) = 1/4$. □

3.32. Algorithm for determining the importance of pages. Internet browsers can find on the Internet (almost) all pages containing a given word or phrase. But how to sort the pages such that the user receives a list 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

dimension of the space V. The operator of derivation on polynomials, $D(x^k) = kx^{k-1}$, is an example of cyclic mapping on spaces $\mathbb{K}_n[x]$ of all polynomials of degree at most n over scalars $\mathbb{K}$.

Surprisingly, this also holds the other way – every nilpotent mapping is a direct sum of cyclic mappings. A proof of this claim takes a lot of work, thus we first formulate the results we are aiming at, and then gradually start with the technical work. In the resulting theorem about *Jordan decomposition* appear vector (sub)spaces and linear mappings on them with a single eigenvalue λ and a matrix

$$J = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & \lambda \end{pmatrix}.$$

These matrices (and corresponding invariant subspaces) are called *Jordan blocks*.

Theorem (Jordan theorem about canonical form)**.** *Let V be a vector space of the dimension n and $f : V \to V$ be a linear mapping with n eigenvalues, counting algebraic multiplicities. Then there exists a unique decomposition of the space V into a direct sum of subspaces*

$$V = V_1 \oplus \dots \oplus V_k$$

such that $f(V_i) \subset V_i$, restriction of f on every V_i has a single eigenvalue λ_i and the restriction $f - \lambda_i \cdot \mathrm{id}$ on V_i is either cyclic or zero mapping.

The theorem thus says that for a suitable basis every linear mapping has block-diagonal form with Jordan blocks along the diagonal. The total number of ones over the diagonal in such form equals the difference between total algebraic and geometric multiplicity of the eigenvalues.

3.33. Notes. Note that we have already proven the Jordan theorem for the cases when all eigenvalues are either distinct or when the geometric and algebraic multiplicities of the eigenvalues are the same. Specifically, we have already proven it for unitary, normal and self-adjoint mappings.

Another useful observation is that for every linear mapping f, every eigenvalue of f has uniquely determined invariant subspace that corresponds to the Jordan block in the matrix.

We should also mention one very useful corollary of the Jordan theorem (which we have already used in the discussion about the behaviour of Markov chains). Assume that the eigenvalues of our mapping f are all in absolute value smaller than one. Then repeated application of the linear mapping on every vector $v \in V$ leads to a fast decrease of all coordinates of $f^k(v)$ bellow any bounds. Indeed, assume for simplicity that on whole V the mapping f has only one eigenvalue λ and $f - \lambda\,\mathrm{id}_V$ is cyclic (that is, we consider only one Jordan block) and let $v_1, \dots, v_\ell$ be the corresponding basis. Then the condition from the theorem says $f(v_2) = \lambda v_2 + v_1$, $f^2(v_2) = \lambda^2 v_2 + \lambda v_1 + \lambda v_1$, and similarly for other v_i and higher powers. In any case, iteration results in higher and higher powers of λ at all non-zero components, while the smallest of them can be at most the degree of nilpotency lower than the number of iterations.

This proves the claim (and the same argument can be used to prove that for the mapping with all eigenvalues with absolute pages as 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 A to page B, determines the probability (1/(total number of links from the page A)), with which the process moves from the page A to the 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 us probabilistic matrix M (the element m_{ij} corresponds to the probability with which we move form 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 point in time (distant enough 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 the users stops clicking from a link to link after certain time and again starts on a random page. Say that with probability d he chooses randomly a new page and with probability $(1-d)$ keeps clicking. In such 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 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. Let us say 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 pages is thus given according to the order of the sizes of the corresponding components, that is, $C > A > B > D$.

3.33. 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 in half of the cases the next day is warm in 50 % of the cases and average day in 30 % of the cases, after an average day the next day is in 40 % of the cases average and cold in 30 % of the cases, and after

value strictly greater than one leads to unbounded growth of all coordinates for the iteration $f^k(v)$).

The rest of this part of the third chapter is devoted to the proof of the Jordan theorem and some necessary lemmata. It is way more difficult than anything so far and the reader can skip it, until the beginning of the fifth part of this chapter.

3.34. Root spaces. On examples we have already seen that the eigensubspaces describe additional geometric properties only for some linear mappings. Thus we now introduce a more subtle tool, the so-called root subspaces.

Definition. Non-zero vector $u \in V$ is called *root vector* of a linear mapping $\varphi : V \to V$, if there exists $a \in \mathbb{K}$ and an integer $k > 0$ such that $(\varphi - a \cdot \mathrm{id}_V)^k(u) = 0$, that is, k-th iteration of the given mapping maps 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 to the scalar $\lambda \in \mathbb{K}$, and is denote as $\mathcal{R}_\lambda$.

If u is a root vector and the k from the definition is chosen the smallest possible, then $(\varphi - a \cdot \mathrm{id}_V)^{k-1}(u)$ 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. *For linear mapping $\varphi : V \to V$ we have:*

(1) for every $\lambda \in \mathbb{K}$ is $\mathcal{R}_\lambda \subset V$ a vector subspace,
(2) for every $\lambda, \mu \in \mathbb{K}$ is $\mathcal{R}_\lambda$ invariant with respect to the linear mapping $(\varphi - \mu \cdot \mathrm{id}_V)$, notably it is that $\mathcal{R}_\lambda$ is invariant with respect to φ,
(3) if $\mu \neq \lambda$, then $(\varphi - \mu \cdot \mathrm{id}_V)|_{\mathcal{R}_\lambda}$ is invertible,
(4) the mapping $(\varphi - \lambda \cdot \mathrm{id}_V)|_{\mathcal{R}_\lambda}$ is nilpotent.

Proof. (1) Checking the properties of the vector vector subspace is easy and we leave it on the reader.

(2) Assume that $(\varphi - \lambda \cdot \mathrm{id}_V)^k(u) = 0$ and consider $v = (\varphi - \mu \cdot \mathrm{id}_V)(u)$. Then

$$\begin{aligned}(\varphi - \lambda \cdot \mathrm{id}_V)^k(v) &= \\ &= (\varphi - \lambda \cdot \mathrm{id}_V)^k((\varphi - \lambda \cdot \mathrm{id}_V) + (\lambda - \mu) \cdot \mathrm{id}_V)(u) \\ &= (\varphi - \lambda \cdot \mathrm{id}_V)^{k+1}(u) + (\lambda - \mu) \cdot (\varphi - \lambda \cdot \mathrm{id}_V)^k(u) \\ &= 0\end{aligned}$$

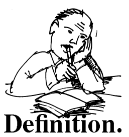
(3) If $u \in \mathrm{Ker}(\varphi - \mu \cdot \mathrm{id}_V)|_{\mathcal{R}_\lambda}$, then

$$(\varphi - \lambda \cdot \mathrm{id}_V)(u) = (\varphi - \mu \cdot \mathrm{id}_V)(u) + (\mu - \lambda) \cdot u = (\mu - \lambda) \cdot u$$

>From there we have that $0 = (\varphi - \lambda \cdot \mathrm{id}_V)^k(u) = (\mu - \lambda)^k \cdot u$ and thus also $u = 0$ for $\lambda \neq \mu$.

(4) Choose a basis $e_1, \dots, e_p$ of the subspace $\mathcal{R}_\lambda$. Because according to the definition there exist numbers k_i such that $(\varphi - \lambda \cdot \mathrm{id}_V)^{k_i}(e_i) = 0$, we have that the whole mapping $(\varphi - \lambda \cdot \mathrm{id}_V)|_{\mathcal{R}_\lambda}$ is nilpotent. □

3.35. Factor subspaces. Our next aim is to show that the dimension of the root spaces is always equal to the algebraic multiplicity of the corresponding eigenvalues. Let us first introduce some useful technical tools.



Definition. Let $U \subset V$ be a vector subspace. On the set of all vectors in V we define an equivalence relation as follows: $v_1 \sim v_2$ 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, along with the operations defined using representants, that is, $[v] + [w] =$

a cold day the next day is in 50 % of the cases cold and in 30 % of the cases average. 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

$$x_{n+1} = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \cdot x_n$$

give the probabilities that the next day is warm, average and cold respectively. For verifying 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 follows a warm, average and cold day (respectively). We see that the problem is a Markov chain 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, thanks to the corollary of the Perron-Frobenius theorem x_∞ is the eigenvector of the matrix T for the eigenvalue 1. Thus it must hold that

$$\begin{aligned} x_\infty^1 &= 0.5\,x_\infty^1 + 0.3\,x_\infty^2 + 0.2\,x_\infty^3, \\ x_\infty^2 &= 0.3\,x_\infty^1 + 0.4\,x_\infty^2 + 0.3\,x_\infty^3, \\ x_\infty^3 &= 0.2\,x_\infty^1 + 0.3\,x_\infty^2 + 0.5\,x_\infty^3, \\ 1 &= x_\infty^1 + x_\infty^2 + x_\infty^3, \end{aligned}$$

where the last condition means that the vector x_∞ is probabilistic. It is easy to compute that this system has a single 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.

Let us emphasise that the sum of the numbers from any column of the matrix T had to equal 1 (otherwise it would not be a Markov process). Because $T^T = T$ (the matrix is symmetric), the sum of all numbers from any row is also equal 1. We say that a matrix with non-negative elements and with the property that the sum of the numbers

$[v + w]$, $a \cdot [u] = [a \cdot u]$, forms a vector space which we call *factor vector space* of the space V by the subspace U.

Check the correctness of the definition of the operations and that all the axiom of the vector space hold!

Classes (vectors) in the factor space V/U will be often denoted as a formal sum of one representant with all vectors of the subspace U, for instance $u + U \in V/U$, $u \in V$. Zero vector is in V/U exactly the class $0 + U$, that is, the vector $u \in V$ represents the zero element in V/U if and only if it is $u \in U$.

For simple examples, think about $V/\{0\} \cong V$, $V/V \cong \{0\}$ and about the factor space of the plane $\mathbb{R}^2$ by any one-dimensional subspace (here, every one-dimensional subspace $U \subset \mathbb{R}^2$ is a line passing through the origin), where the classes of equivalence are parallel lines with this line.

Proposition. *Let $U \subset V$ be a vector subspace and $(u_1, \dots, u_n)$ be such basis of V such that $(u_1, \dots, u_k)$ is a basis of U. Then* $\dim V/U = n - k$ *and the vectors*

$$u_{k+1} + U, \dots, u_n + U$$

form a basis of V/U.

Proof. Because $V = \langle u_1, \dots, u_n \rangle$, it is also that $V/U = \langle u_1 + U, \dots, u_n + U \rangle$. But first k generators are zero, thus $V/U = \langle u_{k+1} + U, \dots, u_n + U \rangle$. Assume that $a_{k+1} \cdot (u_{k+1} + U) + \cdots + a_n \cdot (u_n + U) = (a_{k+1} \cdot u_{k+1} + \cdots + a_n \cdot u_n) + U = 0 \in V/U$. That is equivalent to the belonging of a linear combination of the vectors $u_{k+1}, \dots, u_n$ to the subspace U. Because U is generated by the remaining vectors, the combination is necessary zero, that is, all coefficients a_i are zero. □

3.36. Induced mappings on factor spaces. Assume that $U \subset V$ is an invariant subspace with respect to linear mapping $\varphi : V \to V$ and choose basis $u_1, \dots, u_n$ of the space V such that the first k vectors of this basis is a basis of U. In this basis φ has the block matrix $A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$. Then we are able to 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 $u_{k+1} + U, \dots, u_n + U$ on V/U,*

(2) 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 $a \cdot (\varphi(v) + U) = a \cdot \varphi(v) + U = \varphi(a \cdot v) + U$ (because φ is linear), thus the mapping $\varphi_{V/U}$ is well-defined and linear. Furthermore we have directly from the definition of the matrix of the mapping that the matrix $\varphi_{V/U}$ 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 $|D - \lambda \cdot E|$, while characteristic polynomial of the original mapping φ is $|A - \lambda \cdot E| = |B - \lambda \cdot E||D - \lambda \cdot E|$. □

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* in any column equals one and analogously for rows is called doubly stochastic. Important property of every doubly stochastic primitive matrix (for any dimension – the number of states) is that the corresponding vector x_∞ has all the components identical, that is, after sufficiently many iterations all the states in the corresponding Markov chain are attained with the same frequency. □

3.34. John is used to go 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 uniformly 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 uniformly between medium and long. Assume that he has been running like this for a very long. How often does he choose the short one and how often the long one? What is the probability that he chooses a long one when he picked it a week before?

Solution. Clearly it is a Markov process with three possible states – choices for short, medium and long tack. This order of the states gives a probabilistic transitions 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 realise that for instance the second column corresponds to the choice of the medium track in the previous day, which means that with the probability $1/2$ again a medium track will be chosen (the second row) and with probability $1/2$ a long track will be chosen (the third row). Because we have

$$T^2 = \begin{pmatrix} 1/6 & 1/6 & 1/9 \\ 5/12 & 5/12 & 4/9 \\ 5/12 & 5/12 & 4/9 \end{pmatrix},$$

we can use the corollary of the Perron-Frobenius theorem for Markov chains. It is not difficult to compute that eigenvector corresponding to the eigenvalue 1 and which is probabilistic vector, namely:

$$\left(\frac{1}{7}, \frac{3}{7}, \frac{3}{7}\right)^T.$$

The values $1/7$, $3/7$, $3/7$ then give respectively the probabilities that in a randomly chosen day he choose short, medium and long track.

Let John at a certain day (that is, in time $n \in \mathbb{N}$) choose a long track. This corresponds to the probabilistic vector

$$x_n = (0,\ 0,\ 1)^T.$$

of invariant subspaces $\{0\} = V_0 \subset V_1 \subset \cdots \subset V_n = V$ *with dimensions* $\dim V_i = i$. *Under the basis* $u_1, \dots, u_n$ *of the space* V *such that* $V_i = \langle u_1, \dots, u_i\rangle$ *the mapping* φ *has as its matrix the upper triangular matrix:*

$$\begin{pmatrix} \lambda_1 & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix},$$

where $\lambda_1, \dots, \lambda_n$ *is a sequence of the elements of the spectrum.*

Proof. Construction of the subspaces V_i is done inductively. Let $\lambda_1, \dots, \lambda_n$ be elements in the spectrum of the mapping φ, that means that the characteristic polynomial of the mapping φ is in the form $(\lambda-\lambda_1)\cdot\dots\cdot(\lambda-\lambda_n)$. We choose $V_0 = \{0\}$, $V_1 = \langle u_1\rangle$, where u_1 is an arbitrary eigenvector with eigenvalue λ_1. According to the previous theorem is the characteristic polynomial of the mapping φ_{V/V_1} of the form $(\lambda - \lambda_2)\cdot\dots\cdot(\lambda - \lambda_n)$. Assume that we have already constructed linearly independent vectors $u_1, \dots, u_k$ and invariant subspaces $V_i = \langle u_1 \dots, u_i\rangle$, $i = 1, \dots, k < n$ such that the characteristic polynomial of φ_{V/V_k} is of the form $(\lambda - \lambda_{k+1})\cdot \dots \cdot (\lambda - \lambda_n)$ and $\varphi(u_i) \in (\lambda_i \cdot u_i + V_{i-1})$ for all $i = 1, \dots, k$.

Thus there exists an eigenvector $u_{k+1} + V_k \in V/V_k$ of the mapping φ_{V/V_k} with the eigenvalue la_{k+1}. Consider now the space $V_{k+1} = \langle u_1, \dots, u_{k+1}\rangle$. If the vector u_{k+1} is a linear combination of the vectors $u_1, \dots, u_k$ that means that $u_{k+1}+V_k$ is the zero class in V/V_k, but that is not possible. Thus we have $\dim V_{k+1} = k + 1$. It remains to study the induced mapping $\varphi_{V/V_{k+1}}$. Characteristic polynomial of this mapping is of degree $n - k - 1$ and divides the characteristic polynomial of the mapping φ. But adding the vectors $u_1, \dots, u_{k+1}$ to the basis of V yields a block matrix of the mapping φ with upper triangular submatrix B in the left upper corner and zero in the left lower corner, and the diagonal elements are exactly the scalars $\lambda_1, \dots, \lambda_{k+1}$. Therefore the roots of the characteristic polynomial of the induced mapping have the required properties. □

3.37. Notes. If the decomposition of the whole space V into direct sum of eigensubspaces exists, then there exists a basis of eigensubspaces and the previous theorem actually does not say anything interesting. But its strength is that the only assumption is the existence of $\dim V$ roots of the characteristic polynomial (counting multiplicities). That is ensured whenever the field $\mathbb{K}$ is algebraicly closed, for instance the complex numbers $\mathbb{C}$. A direct corollaries of this are interesting claims about the determinant and the trace of the mapping: they are always the product and the sum of the elements in the spectrum respectively. This can be also used for all real matrices. We can always consider them to be complex, calculate what we need, and because both determinant and the trace are algebraic expressions in terms of the elements of the matrix, the results are exactly the values we wanted.

If we are given a scalar product on a vector space V, we can in every inductive step of the previous proof use the fact that it always holds that $V/V_k \simeq V_k^\perp$ and $V_k^\perp \ni u \mapsto (u + V_k) \in V/V_k$. That means that in every class of the factor V/V_k there exists exactly one vector from $V_k^\perp$. Indeed, the factor space by any subspace in a unitary space has this property – if $u, v \in V_k^\perp$ are in the same class, then their difference belongs to $V_k \cap V_k^\perp$, thus they are equal. Thus we can choose as the representant u_{k+1} of the class (the eigenvector

For the following day it holds that

$$x_{n+1} = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix},$$

and after seven days we have

$$x_{n+7} = T^7 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = T^6 \cdot \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

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 $0.428\,569 \approx 3/7 \doteq 0.428\,571$. □

3.35. The production line is not reliable: individual products differ in quality in a non-neglectible way. Furthermore, 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 the next product is of the class in 60 % of the cases and is of quality I in 20 % of the cases, and after a product of the quality III the next product is of the quality III in 50 % of the cases and in 25 % of the cases it is of quality II. Compute the probability that the 18-th product is of the quality I, if the 16-th product is of quality III.

Solution. Let us first solve the problem without using a Markov chain. The event in question is satisfied by the cases (16-th product is of the class III)

- 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 we easily obtain the result

$$0.375 = 0.2 + 0.05 + 0.125.$$

Now let us view the problem as a Markov process. From the statement we have that to the order of the possible states „product is of

φ_{V/V_k}) choose exactly the vector from $V_k^{\perp}$. This modification leads us to the orthogonal basis with the properties required in the claim about triangulation. Therefore there also exists such orthonormal basis:

Corollary (Schur orthogonal triangulation theorem). *Let $\varphi : V \to V$ be arbitrary linear mapping in a (real or complex) unitary space with $m = \dim V$ eigenvalues (counting multiplicities). Then there exists an orthonormal basis of the space V such that the matrix of φ is under this basis upper triangular with eigenvalues $\lambda_1, \ldots, \lambda_m$ on the diagonal.*

3.38. Theorem. *Let φ $V \to V$ be a linear mapping. The sum of root spaces*

$$\mathcal{R}_{\lambda_1}, \ldots, \mathcal{R}_{\lambda_k}$$

that correspond to distinct eigenvalues $\lambda_1 \ldots, \lambda_k$ is direct. Furthermore, for every eigenvalue λ the dimension of the subspace $\mathcal{R}_\lambda$ equals to the algebraic multiplicity of λ.

Proof. We do the proof by induction over the number k of root spaces. Assume that the theorem holds for less than k spaces and that for vectors $u_1 \in \mathcal{R}_{\lambda_1}, \ldots, u_k \in \mathcal{R}_{\lambda_k}$ we have that $u_1 + \cdots + u_k = 0$. For suitable j then $(\varphi - \lambda_k \cdot \mathrm{id}_V)^j(u_k) = 0$ and also $y_i = (\varphi - \lambda_k \cdot \mathrm{id}_V)^j(u_i)$ are non-zero vectors in $\mathcal{R}_{\lambda_i}$, $i = 1, \ldots, k-1$, if u_i are non-zero (see the previous theorem).

But also

$$y_1 + \cdots + y_{k-1} = \sum_{i=1}^{k} (\varphi - \lambda_k \cdot \mathrm{id}_V)^j(u_i) = 0$$

and thus using the inductive assumption all y_i are non-zero. But then also $u_k = 0$ and linear independence is proven.

It remains to show that the dimension of every root space $\mathcal{R}_\lambda$ equals the algebraic multiplicity of the root λ of the characteristic polynomial. Let thus be λ an eigenvalue of φ, denote by $\bar{\varphi}$ the restriction $\varphi|_{\mathcal{R}_\lambda}$ and let $\psi : V/\mathcal{R}_\lambda \to V/\mathcal{R}_\lambda$ be the mapping induced by φ on the factor space. Assume that the dimension $\mathcal{R}_\lambda$ is smaller than the multiplicity of the root λ of the characteristic polynomial. Using the lemma 3.36 we conclude that 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, $\psi(v+\mathcal{R}_\lambda) = \lambda\cdot(v+\mathcal{R}_\lambda)$, which according to the definition denotes $v \notin \mathcal{R}_\lambda$ and $\varphi(v) = \lambda \cdot v + w$ for suitable $w \in \mathcal{R}_\lambda$. We thus have $w = (\varphi-\lambda\cdot\mathrm{id}_V)(v)$ a $(\varphi-\lambda\cdot\mathrm{id}_V)^j(w) = 0$ for suitable j. We have thus derived $(\varphi-\lambda\cdot\mathrm{id}_V)^{j+1}(v) = 0$, which contradict the choice $v \notin \mathcal{R}_\lambda$.

This proves that the dimension of $\mathcal{R}_\lambda$ equals the multiplicity of the root λ of the characteristic polynomial of φ. □

Corollary. *For every linear mapping $\varphi : V \to V$ whose whole spectrum is in $\mathbb{K}$ is $V = \mathcal{R}_{\lambda_1} \oplus \cdots \oplus \mathcal{R}_{\lambda_n}$ the direct sum of the root subspaces. If we choose suitable bases for these subspaces, then φ has under this basis block-diagonal form with upper triangular matrices in the blocks and eigenvalues λ_i on the diagonal.*

3.39. 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 compose together the already proven results.

class I", „product is of class II", „product is of class III" corresponds the probabilistic matrix

$$\begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix}.$$

Situation that the product is in the class III is given by the probabilistic vector $(0.0.1)^T$. For the next product we obtain the probabilistic vector

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and for the next product in order then the vector

$$\begin{pmatrix} 0.375 \\ 0.3 \\ 0.325 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix},$$

whose first component is the desired probability.

Let us add that the first method of the solution (without using the Markov process) led to the result faster. But let us realise how unclear it would become if we wanted to compute, say, 22-nd or 30-th product. In the second method one can in a sense restrict the computations to 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. □

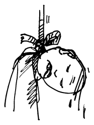
3.36. Repeated dice casting. Write down the transitional probabilistic matrix T for the Markov chain with states „maximum resulting number after n attempts" with the order of the states $1, \dots, 6$. Then determine T^n for every $n \in \mathbb{N}$.

Solution. We can immediately list

$$T = \begin{pmatrix} 1/6 & 0 & 0 & 0 & 0 & 0 \\ 1/6 & 2/6 & 0 & 0 & 0 & 0 \\ 1/6 & 1/6 & 3/6 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 4/6 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 5/6 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1 \end{pmatrix},$$

where 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 $2, \dots, 6$ (the result on the dice would be $2, \dots, 6$), 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 $3, \dots, 6$ (the result would be $3, \dots, 6$), and the last column is derived from the fact that the state 6 is persistent (if 6 has already been seen, no greater result can be).

Theorem. *Let φ $V \to V$ be a nilpotent linear mapping. Then there exists a decomposition of V into a direct sum of subspaces $V = V_1 \oplus \dots \oplus V_k$ such that the restriction of φ on any of them is cyclic.*



Proof. Verifying this is quite straightforward and consists of construction of such basis of the space V that the action of the mapping φ on the basis vectors directly show the decomposition into the cyclic mappings. But taking care of the details will take some time.

Let k be the degree of nilpotency of the mapping φ and denote $P_i = \operatorname{im}(\varphi^i)$, $i = 0, \dots, k$, that is,

$$\{0\} = P_k \subset P_{k-1} \subset \dots \subset P_1 \subset P_0 = V.$$

Choose arbitrary basis $e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1}$ of the space P_{k-1}, where $p_{k-1} > 0$ is the dimension of P_{k-1}. >From the definition it follows that $P_{k-1} \subset \operatorname{Ker} \varphi$, that is, always $\varphi(e_j^{k-1}) = 0$.

Assume that $P_{k-1} \neq V$. Because $P_{k-1} = \varphi(P_{k-2})$, there necessarily exist in P_{k-2} the vectors e_j^{k-2}, $j = 1, \dots, p_{k-1}$ such that $\varphi(e_j^{k-2}) = e_j^{k-1}$. Assume

$$a_1 e_1^{k-1} + \dots + a_{p_{k-1}} e_{p_{k-1}}^{k-1} + b_1 e_1^{k-2} + \dots + b_{p_{k-1}} e_{p_{k-1}}^{k-2} = 0.$$

Application of the mapping φ on this linear combination yields $b_1 e_1^{k-1} + \dots + b_{p_{k-1}} e_{p_{k-1}}^{k-1} = 0$, therefore all $b_j = 0$. But then also $a_j = 0$, because it is a combination of the basis vectors. Thus we have verified the linear independence of all $2p_{k-1}$ chosen vectors. We extend them to a basis

$$\begin{aligned} &e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1} \\ &e_1^{k-2}, \dots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2} \end{aligned}$$

of the space P_{k-2}. Furthermore, the images of the added basis vectors are in P_{k-1}, necessarily they must be linear combinations of the basis elements $e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1}$. We can thus exchange the chosen vectors $e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2}$ with vectors $e_j^{k-2} - \varphi(e_j^{k-2})$. This ensures that the vectors added to the basis of P_{k-2} belong to the kernel of the mapping φ. Let us thus assume it right about the chosen basis (1).

Let us assume further that we have already constructed a basis of the subspace $P_{k-\ell}$ such that we can directly compose it into the schema

$$\begin{aligned} &e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1} \\ &e_1^{k-2}, \dots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2} \\ &e_1^{k-3}, \dots, e_{p_{k-1}}^{k-3}, e_{p_{k-1}+1}^{k-3}, \dots, e_{p_{k-2}}^{k-3}, e_{p_{k-2}+1}^{k-3}, \dots, e_{p_{k-3}}^{k-3} \\ &\vdots \\ &e_1^{k-\ell}, \dots, e_{p_{k-1}}^{k-\ell}, e_{p_{k-1}+1}^{k-\ell}, \dots, e_{p_{k-2}}^{k-\ell}, e_{p_{k-2}+1}^{k-\ell}, \dots, e_{p_{k-3}}^{k-\ell}, \dots e_{p_{k-\ell}}^{k-\ell} \end{aligned}$$

where the value of the mapping φ on any basis vector is located above him, or equals zero if there is nothing above that basis vector. If $P_{k-\ell} \neq V$, then again there must exist vectors $e_1^{k-\ell-1}, \dots, e_{p_{k-\ell}}^{k-\ell-1}$ which map on $e_1^{k-\ell}, \dots, e_{p_{k-\ell}}^{k-\ell}$ and we can extend them to a basis P_{k-l-1}, say by the vectors

$$e_{p_{k-\ell}+1}^{k-\ell-1}, \dots, e_{p_{k-\ell-1}}^{k-\ell-1}.$$

By gradual subtraction of the values through iteration of the mapping φ on these vectors yields that the vectors added to the basis

Also, for $n \in \mathbb{N}$ we can directly determine

$$T^n = \begin{pmatrix} \left(\frac{1}{6}\right)^n & 0 & 0 & 0 & 0 & 0 \\ \left(\frac{2}{6}\right)^n - \left(\frac{1}{6}\right)^n & \left(\frac{2}{6}\right)^n & 0 & 0 & 0 & 0 \\ \left(\frac{3}{6}\right)^n - \left(\frac{2}{6}\right)^n & \left(\frac{3}{6}\right)^n - \left(\frac{2}{6}\right)^n & \left(\frac{3}{6}\right)^n & 0 & 0 & 0 \\ \left(\frac{4}{6}\right)^n - \left(\frac{3}{6}\right)^n & \left(\frac{4}{6}\right)^n - \left(\frac{3}{6}\right)^n & \left(\frac{4}{6}\right)^n - \left(\frac{3}{6}\right)^n & \left(\frac{4}{6}\right)^n & 0 & 0 \\ \left(\frac{5}{6}\right)^n - \left(\frac{4}{6}\right)^n & \left(\frac{5}{6}\right)^n - \left(\frac{4}{6}\right)^n & \left(\frac{5}{6}\right)^n - \left(\frac{4}{6}\right)^n & \left(\frac{5}{6}\right)^n - \left(\frac{4}{6}\right)^n & \left(\frac{5}{6}\right)^n & 0 \\ 1 - \left(\frac{5}{6}\right)^n & 1 - \left(\frac{5}{6}\right)^n & 1 - \left(\frac{5}{6}\right)^n & 1 - \left(\frac{5}{6}\right)^n & 1 - \left(\frac{5}{6}\right)^n & 1 \end{pmatrix}$$

The values in the first column correspond gradually 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 „n-times in a row the result is 1, 2, 3 or 4“, „n-times in a row the result is 1, 2, 3, 4 or 5 and at least once it is 5“ and „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 quick expression of the powers T^n, $n \in \mathbb{N}$. □

3.37. In this problem we deal with a certain property of an animal species which is determined independently of the sex but just by a certain gene – a tuple of alleles. Every individual gains one allele from each of its 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 tuple aA. Exactly one of their offspring (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 individual with the tuple 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 you carry on with this for a long time, compute the probability that both bred individuals have a tuple of alleles AA, or aa (then the process of breeding ends).

$P_{k-\ell-1}$ lie in the kernel of φ and analogically as before we verify that we indeed obtain a basis $P_{k-\ell-1}$.

After k steps we obtain a basis of the whole V, which has the properties given for the basis of the subspace $P_{k-\ell}$. Individual columns of the resulting schema then generate the subspaces V_i and additionally we have directly found the bases of these subspaces that show that corresponding restrictions of φ are cyclic mappings. □

3.40. Proof of the Jordan theorem. Let $\lambda_1, \dots, \lambda_k$ be all 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 $\varphi_i = (\varphi|_{\mathcal{R}_{\lambda_i}} - \lambda_i \cdot \mathrm{id}_{\mathcal{R}_{\lambda_i}})$ 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 $\varphi - \lambda_i \cdot \mathrm{id}_V$ 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 prove the claim about uniqueness. Because the diagonal values λ_i are given as roots of the characteristic polynomial, their uniqueness is immediate. We express the dimensions of individual Jordan blocks using the ranks $r_k(\lambda_i)$ of the mapping $(\varphi - \lambda_i \cdot \mathrm{id}_V)^k$. This will show that the blocks are uniquely determined (up to their order). On the other hand, switching 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 n-dimensional space, then the defect of the iterated mapping ψ^k is k for $0 \le k \le n$ and is n for all $k \ge n$. This implies that if the matrix J of the mapping φ contains $d_k(\lambda)$ of Jordan blocks of the order k with the eigenvalue λ, then the defect of the matrix $(J - \lambda \cdot E)^\ell$ is

$$d_1(\lambda) + 2d_2(\lambda) + \dots \ell d_\ell(\lambda) + \ell d_{\ell+1}(\lambda) + \dots$$

>From here we calculate

$$n - r_\ell(\lambda) = d_1(\lambda) + 2d_2(\lambda) + \cdots + \ell d_\ell(\lambda) + \ell d_{\ell+1}(\lambda) + \dots$$
$$d_k(\lambda) = r_{k-1}(\lambda) - 2r_k(\lambda) + r_{k+1}(\lambda)$$

(where the last row arises by combining the previous for values $\ell = k-1, k, k+1$).

3.41. Note. The proof of the theorem about the existence of the Jordan canonical form was constructive, but it does not give us a perfect algorithmic approach for the construction. Now we summarise the already derived approach for the explicit computation of the basis under which the given mapping $\varphi : V \to V$ has the matrix in the canonical Jordan form.

(1) We find the roots of the characteristic polynomial.
(2) If there are less than $n = \dim V$ of them (counting multiplicities), there is no canonical form.
(3) If there are n linearly independent eigenvectors, we obtain a basis of V composed of eigenvectors and under it φ has diagonal matrix.
(4) Let λ be the eigenvalue with geometric multiplicity strictly smaller than algebraic multiplicity and $v_1, \dots, v_k$ be the corresponding eigenvectors. They should be the vectors on the

(d) Solve the problem from the case (c) without the condition that the individuals have the same parent. Thus you just breed random individuals from a population among them, then you breed among their progeny, and so on.

Solution. Case (a). It is a Markov process given by the matrix

$$T = \begin{pmatrix} 1/2 & 1/4 & 0 \\ 1/2 & 1/2 & 1/2 \\ 0 & 1/4 & 1/2 \end{pmatrix},$$

while the order of the states corresponds to the order of the tuples of alleles aa, aA, AA. The values in the first column follow from the fact that an offspring of parents that have alleles aa and aA has probability $1/2$ for the tuple aa and probability $1/2$ for the tuple aA. Analogously, we work out the third column. The values in the second column follow from the fact that each of the four cases of the tuples of alleles aa, aA, Aa, AA has the same probability for an individual whose both parents have the tuple 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 tuples aA and Aa (which are then the same). For determining the resulting state it thus then suffices 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 (all 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 appearance of combinations aa, aA and AA respectively, after a very long (theoretically infinite) time.

Case (b). For the order of the tuples of alleles AA, aA, aa we now obtain the probabilistic matrix

$$T = \begin{pmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

We immediately see all eigenvalues 1, $1/2$ and 0 (if we subtract them from the diagonal, the rank of the resulting matrix is not 3, that is, the homogeneous system given by this matrix will have a non-trivial solution). To these eigenvectors then respectively correspond the eigenvectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

upper border of the scheme from the proof of the theorem 3.39, but it is necessary to find a suitable basis by application of iterations $\varphi - \lambda \cdot \mathrm{id}_V$. By doing this we also find out in which row are the vectors located, and we find the linearly independent solutions of the equations $(\varphi - \lambda\,\mathrm{id})(x) = v_i$ from the rows bellow it. We repeat the procedure iteratively (that is, for w_i and so no). We find by this "chains" of basis vectors that give subspaces, where $\varphi - \lambda\,\mathrm{id}$ are cyclic.

The procedure is practical for matrices where the multiplicities of the eigenvalues are small, or at least 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

$$\langle (1, 0, 0), (0, 1, 0) \rangle.$$

We need to find the solutions of the equations $(A - 2E)x = (a, b, 0)^T$ for suitable constants a, b. This system is solvable only for $a = b$ and one of the possible solutions is $v = (0, 0, 1)$, $a = b = 1$. The whole basis is then composed of $(1, 1, 0)$, $(0, 0, 1)$, $(1, 0, 0)$. Note that we had many choices for bases and thus there are many such bases.

5. Decompositions of the matrices and pseudoinversions

In the previous part we concentrated on the geometric description of the structure of the mapping. Now we translate our results into the language of the so-called matrix decompositions, which is a very important topic for numerical methods and matrix calculus in general.

Even when computing with real numbers we are using for simplicity decompositions into products. The simplest is expression of every real number uniquely in the form

$$a = \mathrm{sgn}(a) \cdot |a|,$$

that is, as a product of the sign and the absolute value. In the following text we briefly list some of such decompositions for distinct types of matrices. For instance, we have already used a suitable decomposition for positively semidefinite matrices in the paragraph 3.31 for construction of the square root of the matrix.

3.42. LU-decomposition. Let us begin with reformulation of some results we have already derived. In the paragraphs 2.7 and 2.8 we have transformed matrices over scalars from any field into the row echelon form. For this we have used elementary row transforms, which were based on gradual multiplication of our matrix by invertible lower triangular matrices P_i which acted by adding multiples of rows under the one transformed at the moment.

Assume for simplicity that our matrix A is square and that Gaussian elimination does not force us to swap rows – thus all our matrices P_i can be lower triangular with ones on diagonal. Finally, it suffices to note that inverses of such P_i are again lower triangular with ones on the diagonal and we obtain

$$U = P \cdot A = P_k \cdots P_1 \cdot A$$

Therefore it is

$$T = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

>From there for arbitrary $n \in \mathbb{N}$ it follows

$$T^n = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}^n \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2^{-n} & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Clearly for big $n \in \mathbb{N}$ we can substitute 0 for 2^{-n}, which implies

$$T^n \approx \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Thus if individuals of the original population procreate exclusively with the member of the specific population (that, that has only AA), necessarily after a sufficient number of breeding it results into a total elimination of the tuples aA and aa (and it does not matter what their original distribution was).

The case (c). Now we have 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 tuple AA and the second has aA, then clearly each of the four cases (we are talking about the tuple 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 $1/4$.

Now we should again determine the powers T^n for big $n \in \mathbb{N}$. Considering the form of the first and of the last column we immediately

where U is the upper triangular matrix and 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.

In the case of the general matrix we can with Gaussian elimination into the row echelon form need some additional row permutations, sometimes even column permutations. Then we obtain the more general

$$A = P \cdot L \cdot U \cdot Q,$$

where P and Q are some permutation matrices.

3.43. Notes. A direct corollary of the Gaussian elimination is also a realisation that up to the choice of suitable bases on the domain and codomain, every mapping $f : V \to W$ given by a matrix in block-diagonal form with unit matrix, with size given by the dimension of the image 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.

For square matrices we have in 3.32 shown when discussing properties of linear mappings $f : V \to V$ over complex vector spaces that every square matrix A of the dimension m can be decomposed into the product

$$A = P \cdot B \cdot P^{-1},$$

where B is block-diagonal with Jordan blocks associated to eigenvalues on the diagonal. Indeed, it 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 a change of the basis on the vector space V and the cited theorem says that in a suitable basis every mapping has Jordan canonical form.

Analogously, when discussing the self-adjoint mappings we have proved that for real symmetric matrices or for complex Hermitian matrices there always exists a decomposition into the product

$$A = P \cdot B \cdot P^*,$$

where B is a diagonal matrix with all (always real) eigenvalues on the diagonal, counting multiplicities. Indeed, it is again a product of matrices then stand for the change of the basis, but we allow only changes between orthonormal bases and thus also the matrix P for the change must be orthogonal. From there we have $P^{-1} = P^*$.

For real orthogonal mappings we have derived analogous expression as for symmetric, only our B is diagonal with blocks of size two or one, expressing either rotation or mirror symmetry or identity with respect to the corresponding subspaces.

3.44. Singular decomposition theorem. Let us return to general linear mappings between vector spaces (in general distinct). If a scalar product is defined on them and we restrict ourselves on orthonormal bases only, we must proceed in a more refined way than in the case of arbitrary bases.

find out that 1 is an eigenvalue of the matrix T. It is very easy to find the eigenvectors

$$(1, 0, 0, 0, 0, 0)^T, \quad (0, 0, 0, 0, 0, 1)^T$$

corresponding to the eigenvalue 1. By considering only a four-dimensional submatrix of the matrix T (omitting the first and sixth row and column) we find the remaining eigenvalues

$$\frac{1}{2}, \quad \frac{1}{4}, \quad \frac{1-\sqrt{5}}{4}, \quad \frac{1+\sqrt{5}}{4}.$$

If we recall the solution of the exercise called Sweet-toothed gambler, we don't 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 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 only at the first and sixth position of the resulting vector can be a non-zero number, the states

$$aA, AA; \quad aa, AA; \quad aA, aA; \quad aa, aA$$

after many breedings disappear. Let us further realise (follows also from the exercise Sweet-toothed gambler) that the probability that the process ends with AA, AA equals the relative ratio of the appearance of A in the initial state.

The case (d). Let the values $a, b, c \in [0, 1]$ give in this order the relative ratios of occurrence of alleles AA, aA, aa in the given population. We want to obtain the expression of relative ratios of the tuples AA, aA, aa in the offspring of the population. If the choice of tuples for breeding is random, then for a suitably big population it can be expected that the relative ratio of breeding of individuals that both have AA is a^2, the relative ratio for the tuple aA and AA is $2ab$, the relative ratio for aA (both of them) is b^2 and so on. The offspring of the parents with tuples AA, AA must inherit AA. The probability that the offspring of the parents with tuples AA, aA has AA is clearly $1/2$ and the probability that the offspring of the parents with tuples aA, aA has AA is $1/4$. There are no other cases for an offspring with the tuple AA (if one of the parents has the tuple aa, then the offspring cannot have AA). Relative frequency of AA in the progeny is thus

$$a^2 \cdot 1 + 2ab \cdot \frac{1}{2} + b^2 \cdot \frac{1}{4} = a^2 + ab + \frac{b^2}{4}.$$

Theorem. *Let A be any 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*

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

and r is the rank of the matrix AA^*. Furthermore, S is determined uniquely up to the order of the elements and the elements of the diagonal matrix D are the square roots of the eigenvalues d_i of the matrix AA^*. If A is a real matrix, then the matrices U and V are orthogonal.

Proof. Assume first that $m \leq n$ and denote $\varphi : \mathbb{K}^n \to \mathbb{K}^m$ the mapping between real and complex spaces with standard scalar products, given by the matrix A under the standard bases.

We can reformulate the statement of the theorem as follows: there exists orthonormal bases on $\mathbb{K}^n$ and $\mathbb{K}^m$ under which the mapping φ has the matrix S from the claim of the theorem.

As we have seen before, the matrix A^*A is positively semidefinite. Therefore it has only real non-negative eigenvalues and there exists an orthonormal basis $\underline{w}$ in $\mathbb{K}^n$ under which the corresponding mapping $\varphi^* \circ \varphi$ has for matrix a diagonal matrix with eigenvalues on the diagonal. In other words, there exists unitary matrix V such that $A^*A = VBV^*$ for real diagonal matrix with non-negative eigenvalues $(d_1, d_2, \dots, d_r, 0, \dots, 0)$ on the diagonal, $d_i \neq 0$ for all $i = 1, \dots, r$. >From there

$$B = V^*A^*AV = (AV)^*(AV).$$

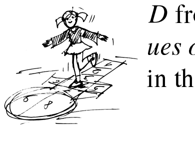
That is equivalent to the claim that first r columns of the matrix AV are orthogonal and the remaining are zero, because they have zero size.

Let us now denote first r columns $v_1, \dots, v_r \in \mathbb{R}^m$. Thus it holds that $\langle v_i, v_i \rangle = d_i$, $i = 1, \dots, r$, and the normalised vectors $u_i = \frac{1}{\sqrt{d_i}} v_i$ form an orthonormal system of non-zero vectors. Let us extend them to an orthonormal basis $\underline{u} = u_1, \dots, u_n$ of the whole $\mathbb{K}^m$. If we express our original mapping φ under the bases $\underline{w}$ in $\mathbb{K}^n$ and $\underline{u}$ in $\mathbb{K}^m$, we obtain the matrix $\sqrt{B}$. The transformations from the standard bases to the new chosen ones correspond to the multiplication from the left with orthogonal matrix U and from the right with $V^{-1} = V^*$.

If $m > n$, we can apply the previous part of the proof on the matrix A^*. From there we directly obtain the desired claim.

If we work over real scalars, all the previous steps in the proof are also realised in the real domain. □

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.



3.45. Geometric interpretation. Diagonal values of the matrix D from the previous theorem are called *singular values of the matrix A*. Let us reformulate this theorem in the real case more geometrically.

Analogically we set gradually the relative frequencies of the tuples 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 $(a, b, c)^T$. It holds that

$$T : \begin{pmatrix} a \\ b \\ c \end{pmatrix} \mapsto \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix}.$$

Let us mention that the domain (and also the codomain) of T are just the vectors

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \text{kde} \quad a, b, c \in [0, 1],\ a + b + c = 1.$$

We would like to give the operation T my multiplying the vector by some constant matrix. But that is clearly not possible (the mapping T is not linear). It is thus 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 compute what happens if we apply the mapping T twice in a row. In the second step we obtain

$$T : \begin{pmatrix} a^2 + ab + b^2/4 \\ ab + bc + 2ac + b^2/2 \\ c^2 + bc + b^2/4 \end{pmatrix} \mapsto \begin{pmatrix} t_2^1 \\ t_2^2 \\ t_2^3 \end{pmatrix}, \quad \text{kde}$$

$$t_2^1 = \left(a^2 + ab + \frac{b^2}{4}\right)^2 + \left(a^2 + ab + \frac{b^2}{4}\right)\left(ab + bc + 2ac + \frac{b^2}{2}\right) + \frac{1}{4}\left(ab + bc + 2ac + \frac{b^2}{2}\right)^2,$$

$$t_2^2 = \left(a^2 + ab + \frac{b^2}{4}\right)\left(ab + bc + 2ac + \frac{b^2}{2}\right) + \\ + \left(ab + bc + 2ac + \frac{b^2}{2}\right)\left(c^2 + bc + \frac{b^2}{4}\right) + \\ + 2\left(a^2 + ab + \frac{b^2}{4}\right)\left(c^2 + bc + \frac{b^2}{4}\right) + \frac{1}{2}\left(ab + bc + 2ac + \frac{b^2}{2}\right)^2,$$

$$t_2^3 = \left(c^2 + bc + \frac{b^2}{4}\right)^2 + \left(ab + bc + 2ac + \frac{b^2}{2}\right)\left(c^2 + bc + \frac{b^2}{4}\right) + \frac{1}{4}\left(ab + bc + 2ac + \frac{b^2}{2}\right)^2.$$

It can be shown (using $a + b + c = 1$) that

$$t_2^1 = a^2 + ab + \frac{b^2}{4}, \quad t_2^2 = ab + bc + 2ac + \frac{b^2}{2}, \quad t_2^3 = c^2 + bc + \frac{b^2}{4},$$

For the corresponding linear mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ the singular values have indeed simple geometric meaning: let $K \subset \mathbb{R}^n$ be the unit ball for the standard scalar product. The image $\varphi(K)$ is the always an m-dimensional ellipsoid (possibly degenerate). The singular values of the matrix A are then the sizes of the main half-axes and the theorem further says that the original sphere always allows orthogonal grouped diameters, whose image are exactly all 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 the sequence of numerical computations with matrices, for instance for the computation of the inverse matrix. Let us also note that there exist fast methods of computations (approximations) for eigenvalues, thus the singular decomposition is very effective to work with.

3.46. Polar decomposition theorem. The singular decomposition theorem is a starting point for many very useful tools. Let us now think about some direct corollaries (which by themselves are quite non-trivial). The statement of the theorem says 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. But then also $A = USU^*UW^*$ and let us call the matrices $P = USU^*$, $V = UW^*$. First of them, P, is Hermitian (in real case symmetric) and positively semidefinite, because it regards just how to write down the mapping with real diagonal matrix S in another orthonormal basis, while V is a product of two unitary matrices and thus again unitary (in the real case orthogonal). Furthermore $A^* = WSU^*$ and thus $AA^* = USSU^* = P^2$ and our matrix P is actually the square root of the easily computable Hermitian matrix AA^*.

Assume that $A = PV = QU$ are two such decompositions of the matrix A into the product of positively semidefinite Hermitian and unitary matrix and assume that A is invertible. But then

$$AA^* = PVV^*P = P^2 = QUU^*Q = Q^2$$

is positively definite and thus the matrices $Q = P = \sqrt{AA^*}$ are uniquely determined and invertible. But then also $U = V = P^{-1}A$.

We have thus completely derived a very useful analogy of the decomposition of a real number into a sign (orthogonal matrix in the case of dimension are exactly ±1) and the absolute value (the matrix P, for which we can compute the square root).

Theorem (Polar decomposition theorem)**.** *Every square complex matrix A of the dimension n can be always expressed in the form $A = P \cdot V$, where P is Hermitian and positively definite square matrix of the same dimension and V is unitary. We have $P = \sqrt{AA^*}$. If A is invertible, the decomposition is unique and $V = (\sqrt{AA^*})^{-1}A$.*

If we work over real scalars, P is symmetric and V orthogonal.

If we apply the same theorem on A^* instead of A, we obtain the same result, but with the order of the Hermitian and unitary matrices reversed. The matrices in the corresponding right and left decomposition will of course be in general distinct.

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 obtain a surprising result that further application of the transform T does not change the vector obtained in the first step. That means that the appearance of the considered tuples is after arbitrary long time time the same as in the first generation of offspring. For a big population we have thus proven that the evolution takes place during first generation (unless there are some mutation or selection). □

3.38. Let there are two boxes, which contain together n white and n black balls. In regular time intervals from both boxes a ball is taken and moved to the other urn, while the number of balls in each of the boxes is at the beginning (and thus for all the time) equal to n. Give for this Markov process its probabilistic transition matrix T.

Solution. This case is often used in physics as a model of blending two incompressible liquids (already in the year 1769 introduced by D. Bernoulli) or analogously, as a model of diffusion of gases. The states $0, 1, \dots, n$ correspond for instance to the number of white balls in the first box. This information already says how many black balls are in the first box (and the remaining balls are then in the second box). If in the certain step a state changes from $j \in \{1, \dots, n\}$ to $j-1$, it means that from the first box a white ball was drawn and from the second a black ball was drawn. That happens with probability

$$\frac{j}{n}\cdot\frac{j}{n}=\frac{j^2}{n^2}.$$

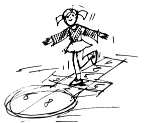
Transition from the state $j \in \{0, \dots, n-1\}$ to the state $j+1$ corresponds to drawing the black ball from the first box and a white ball from the second box, with probability

$$\frac{n-j}{n}\cdot\frac{n-j}{n}=\frac{(n-j)^2}{n^2}.$$

The system stays in the state $j \in \{1, \dots, n-1\}$, if from both boxes balls of the same colour were drawn, which has the same probability

$$\frac{j}{n}\cdot\frac{n-j}{n}+\frac{n-j}{n}\cdot\frac{j}{n}=\frac{2j\,(n-j)}{n^2}.$$

Let us add 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



In the complex case the analogy with the decomposition of numbers is even more funny – positively semidefinite P again plays a role of the absolute value of the complex number, the unitary matrix V then has a unique expression as a sum $V = \text{re}\, V + i\ \text{im}\, V$ with Hermitian real and imaginary parts and the property $(\text{re}\, V)^2 + (\text{im}\, V)^2 = E$, that is, we obtain a full analogy for the polar form for the complex numbers (see the final remark in 3.30). But note that in the case with more dimensions it is important in what order is this "polar form" of matrix written. It is possible in both ways, but the results are in general distinct.

For many practical applications it is faster to use the so-called *QR decomposition* of matrices, which is an analogy of the Schur orthogonal triangulation theorem:

3.47. 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 $A = Q^T R$.*

If we work over real scalars, both Q and R are real.

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 under the standard bases we can choose new orthonormal basis on $\mathbb{K}^m$ such that then 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, and choose from them 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, and we 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 Gramm-Schmidt orthogonalisation of this system of vectors.

Now for every e_i is $\varphi(e_i)$ 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 $\varphi(e_i)$ under the basis $\underline{u}$ appear only vectors $u_1, \dots, u_i$. The mapping φ thus has under the standard basis on $\mathbb{K}^n$ and under $\underline{u}$ on $\mathbb{K}^m$ upper triangular matrix R. The change of the basis $\underline{u}$ on $\mathbb{R}^m$ corresponds to the multiplication by unitary matrix Q from the left, that is, $R = QA$, equivalently $A = Q^T R$.

The last claim is clear from our construction. □

To close this part of the text let us note the especially useful and important application of our results for the approximate numerical calculations. It is a quite straightforward application of singular decompositions of matrices, as can be already seen from the following:

3.48. Definition. Let A be a real matrix of the type m/n and let

$$A = USV^*, \quad S=\begin{pmatrix} D & 0\\ 0 & 0\end{pmatrix}$$

be its singular decomposition (notably, D is invertible). The matrix

$$A^\dagger := VS'U^*, \quad S'=\begin{pmatrix} D^{-1} & 0\\ 0 & 0\end{pmatrix}$$

is called the *pseudoinverse matrix* of the matrix A.

As the following theorem shows, the pseudoinverse is an important generalisation of the notion of inverse matrix, together with direct applications.

3.49. Theorem. *Let A be real or complex matrix of the type m/n. Then for its pseudoinverse it holds that:*

the state $n-1$. Considering it all we obtain the matrix

$$T=\frac{1}{n^2}\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 0\\ n^2 & 2\cdot 1(n-1) & 2^2 & \ddots & 0 & 0 & 0\\ 0 & (n-1)^2 & 2\cdot 2(n-2) & \ddots & 0 & 0 & 0\\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & 0 & \ddots & 2\cdot(n-2)2 & (n-1)^2 & 0\\ 0 & 0 & 0 & \ddots & 2^2 & 2\cdot(n-1)1 & n^2\\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

for the order of the states $0, 1, \dots, n$.

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 certain time in a very negligible way.

If we had numbered the individual balls, we would instead of ball drawing draw some of the numbers $1, 2, \dots, 2n$ and the ball whose number was draw would move to the other ball. We would obtain a Markov process with states $0, 1, \dots, 2n$ (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 have introduced it in 1907). It is used as a model of interchange of heat between two isolated bodies (the heat is represented by the number of balls, the bodies by the boxes). □

3.39. Two players, A and B, gamble for money repeatedly a certain game, which can result only in a victory of one of the players. The winning probability for the player A is in each individual game $p \in [0, 1/2)$ and both bet always only €1, that is, after each game with probability p the player B gives €1 to the player A and with probability $1-p$ the other way round. They play as long as both have some money. If the player A has at the start of the game €x and the player B has €y, determine the probability that the player A loses it all.

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

$$\frac{1-\left(\frac{p}{1-p}\right)^{y}}{1-\left(\frac{p}{1-p}\right)^{x+y}}. \tag{3.6}$$

Let us investigate what is this value for specific choices of p, x, y. If the player B wants to be almost sure and requires that the probability that the player A loses with him €1 000 000 c is at least 0.999, then it

(1) if A is invertible (notably, it is square), then

$$A^{\dagger}=A^{-1},$$

(2) for pseudoinverse $A^{\dagger}$ it holds that $A^{\dagger}A$ and $AA^{\dagger}$ are Hermitian (in real case symmetric) and

$$AA^{\dagger}A=A, \quad A^{\dagger}AA^{\dagger}=A^{\dagger},$$

(3) pseudoinverse matrices $A^{\dagger}$ is by the four properties from the previous point determined uniquely. Thus if some matrix B of the type $n\times m$ satisfies 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 the right-hand side $b\in\mathbb{K}^m$, then the vector $y=A^{\dagger}b\in\mathbb{K}^n$ minimises the size $\|Ax-b\|$ 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 it holds that $AA^{\dagger}b=b$. 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 right from the definition we have $S'=S^{-1}$. From that it follows that $A^{(-1)}A=AA^{(-1)}=E$.

(2): Direct computation yields $SS'S=S$ and $S'SS'=S'$, therefore

$$AA^{(-1)}A=USV^*VS'U^*USV^*=USS'SV^*=USV^*=A$$

and analogically for the second equation. Furthermore,

$$\begin{aligned}(AA^{(-1)})^*&=(USS'U^*)^*=U(S')^*S^*U^*\\&=U(SS')^*U^*=USS'U^*=AA^{(-1)}\end{aligned}$$

and similarly it can be proven that $(A^{(-1)}A)^*=A^{(-1)}A$.

(3) The claim can be proven via direct computation. We consider for a while the mapping φ given under the standard basis by the matrix A, and we express the mapping φ under the basis form the singular decomposition theorem, that is, under this basis the mapping φ has matrix S from the definition of pseudoinverse $A^{\dagger}$. Without loss of generality we now work in this basis, that is, we can assume that in the block form

$$A=\begin{pmatrix}D&0\\0&0\end{pmatrix},\ A^{\dagger}=\begin{pmatrix}D^{-1}&0\\0&0\end{pmatrix},$$

with diagonal matrix D of all non-zero singular numbers, and B is a matrix satisfying the assumptions. Clearly

$$A^{\dagger}A=\begin{pmatrix}E&0\\0&0\end{pmatrix}$$

and thus we obtain

$$A^{\dagger}=A^{\dagger}ABAA^{\dagger}=\begin{pmatrix}E&0\\0&0\end{pmatrix}B\begin{pmatrix}E&0\\0&0\end{pmatrix}=\begin{pmatrix}D^{-1}&0\\0&0\end{pmatrix}.$$

>From there we see that

$$B=\begin{pmatrix}D^{-1}&P\\Q&R\end{pmatrix}$$

for suitable matrices P, Q and R. But now

$$BA=\begin{pmatrix}D^{-1}&P\\Q&R\end{pmatrix}\begin{pmatrix}D&0\\0&0\end{pmatrix}=\begin{pmatrix}E&0\\QD&0\end{pmatrix}$$

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. □

3.40. 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. To the more successful department will then 2 employees of the other department be moved. This process will go on 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 won't 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 estimates that the second one will have relatively greater income in 54 % of the cases? ○

Another application of the Markov chains are in the additional exercises after this chapter.

E. Unitary spaces

Even in the previous chapter we have defined the scalar product in real vector spaces (2.40), in this chapter we extend its definition to the complex spaces too (3.23).

3.41. 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 with respect to the operation of composition a group (see 1.1); composition of two such mappings is by the the definition also a mapping that preserves lengths and angles, unit element of the group is the identity mapping, the inverse element for a given mapping is its inverse mapping – thanks to the condition on the lengths preservation such mapping exists. 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 all invertible mappings from $\mathbb{R}^n$ to $\mathbb{R}^n$.

If we additionally require that the matrices have determinant one, 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 group of all (complex) matrices that correspond to the complex linear mappings from

should be Hermitian, thus $QD = O$ and thus also $Q = O$ (the matrix D is diagonal and invertible). Analogously, the assumption that AB is Hermitian implies that P is zero. Additionally, we have

$$B = BAB = \begin{pmatrix} D^{-1} & 0 \\ 0 & R \end{pmatrix}\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} D^{-1} & 0 \\ 0 & R \end{pmatrix}.$$

On the right side in the right-lower corner there is zero, and thus also $R = O$ and the claim is proven.

(4): Consider the mapping $\varphi : \mathbb{K}^n \to \mathbb{K}^m$, $x \mapsto Ax$, and direct sums $\mathbb{K}^n = (\operatorname{Ker}\varphi)^\perp \oplus \operatorname{Ker}\varphi$, $\mathbb{K}^m = \operatorname{Im}\varphi \oplus (\operatorname{Im}\varphi)^\perp$. The restricted mapping $\tilde{\varphi} := \varphi_{|(\operatorname{Ker}\varphi)^\perp} : (\operatorname{Ker}\varphi)^\perp \to \operatorname{Im}\varphi$ is a linear isomorphism. If we choose suitable orthonormal bases on $(\operatorname{Ker}\varphi)^\perp$ and $\operatorname{Im}\varphi$ 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. For given $b \in \mathbb{K}^m$ is the point $z \in \operatorname{Im}\varphi$ that minimises the distance $\|b - z\|$ (that is, the point that realises the distance from the affine subspace $\rho(b, \operatorname{Im}\varphi)$, see the next chapter) exactly the component $z = b_1$ of the decomposition $b = b_1 + b_2$, $b_1 \in \operatorname{Im}\varphi$, $b_2 \in (\operatorname{Im}\varphi)^\perp$. But in a suitably chosen basis is the mapping $\varphi^{(-1)}$, originally given under standard bases by the pseudoinverse $A^{(-1)}$, given by the matrix S' from the singular decomposition theorem, notably we have $\varphi^{(-1)}(\operatorname{Im}\varphi) = (\operatorname{Ker}\varphi)^\perp$ and D^{-1} is the matrix of the restriction $\varphi^{(-1)}_{|\operatorname{Im}\varphi}$ and $\varphi^{(-1)}_{|(\operatorname{Im}\varphi)^\perp}$ is zero. Indeed we have

$$\varphi \circ \varphi^{(-1)}(b) = \varphi(\varphi^{(-1)}(z)) = z$$

and the proof is finished.

(5) Evidently, from the equality $Ax = b$ for a fixed $x \in \mathbb{K}^n$ it follows that

$$b = AA^\dagger Ax = AA^\dagger b.$$

Thus it is a necessary condition. On the other hand, if this condition holds, then we can for the given expression x compute

$$Ax = A(A^\dagger b + (E - A^\dagger A)u) = b + (A - AA^\dagger A)u = b.$$

The rank of the matrix $A - A^\dagger A$ gives the correct size of the image of the corresponding mapping according to the Frobenius theorem about solution of the system of linear equations, and thus we obtain in this way all solutions. □

Remark. It can be also shown that the matrix $A^{(-1)}$ minimises the expression

$$\|AA^{(-1)} - E\|^2$$

that is, the sum of squares of all elements of the given matrix.

>From the point (4) of the previous theorem we obtain that the matrix $AA^\dagger$ is the matrix of the perpendicular projection form the vector space $\mathbb{R}^n$, where n is the number of the rows of the matrix A on the subspace generated by the columns of the matrix A (this interpretation has of course meaning only for matrices that have more rows than columns).

Furthermore, for matrices A whose columns for independent vectors, the expression $(A^T A)^{-1}A^T$ makes sense and it is not hard to verify that this matrix satisfies all properties from (1) and (2) from the previous theorem, thus it is a pseudoinverse of the matrix A.

$\mathbb{C}^n$ to $\mathbb{C}^n$ that 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, determinant of matrix in $U(n)$ can be any complex unit).

3.42. Consider the vector space V of functions $\mathbb{R} \to \mathbb{C}$. Determine whether the mapping φ from the unitary space V is linear.

i) $\varphi(u) = \lambda u$ where $\lambda \in \mathbb{C}$

ii) $\varphi(u) = u^*$

iii) $\varphi(u) = u^2 (= u.u)$

iv) $\varphi(u) = \frac{du}{dx}$

V is for suitable functions a unitary space of infinite dimension. Scalar product is defined by the relation $f.g = \int_{-\infty}^{\infty} f(x)\overline{g(x)}\,dx$.

Solution. yes, no, no, yes □

3.43. 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.

Solution. >From the definition of exp we can show that it holds that $\exp(A+B) = \exp(A).\exp(B)$ as we are used to with the exponential mapping in the domain of numbers. Because in general it is $(u+v)^* = u^* + v^*$ 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 because $H^* = H$, then

$$U^* = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!}(iH)^n = \exp(-iH)$$

and thus

$$U^*U = \exp(iH)\exp(-iH) = \exp(0) = 1.$$

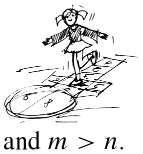
$$\det(U) = \mathrm{e}^{\mathrm{trace}(iH)}\,.$$

□

3.44. 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. We assume that all eigensubspaces of the operator C have dim $= 1$. Then we can for any vector u 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 it clearly holds that

$$0 = [A, C]u_k = ACu_k - CAu_k = \lambda_k Au_k - C(Au_k)$$



3.50. Linear regression. The approximation property (3) from the previous theorem is very useful in the cases where we are to find as good 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 us many measured real values b_j and we want to find a linear combination of some functions f_i, which approximates the values b_j. The actual values of the chose functions in the points $y_j \in \mathbb{R}$ give a matrix $a_{ij} = f_j(y_i)$, whose columns are given by values of the individual functions f_j in the considered points, and our goal is to determine the coefficients $x_j \in \mathbb{R}$ so that the sum of the squares of the deviations from the actual values

$$\sum_{i=1}^{m}(b_i - (\sum_{j=1}^{n} x_j f_j(y_i)))^2 = \sum_{i=1}^{m}(b_i - (\sum_{j=1}^{n} a_{ij}x_j))^2$$

is minimised. In other words, we seek a linear combination of the functions f_i such that we interpolate the given values b_i "well". Thanks to the previous theorem are the optimal coefficients $A^{(-1)}b$.

In order to have a more specific idea, consider just two functions $f_1(x) = x$, $f_2(x) = x^2$ and assume that the "measured values" of their unknown combination $g(x) = y_1 x + y_2 x^2$ in integral values for x between 1 and 10 are $b^T = (1.44\ 10.64\ 4.48\ 14.56\ 31.12\ 39.20\ 54.88\ 71.28\ 85.92\ 104.16)$. This vector arose by computing the values $x + x^2$ in given points shifted by random values in range ±8. The matrix $A = (b_{ij})$ is in our case equal to

$$A^T = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 4 & 9 & 16 & 25 & 36 & 49 & 64 & 81 & 100 \end{pmatrix}$$

and the coefficients in the combination are

$$y = A^{(-1)} \cdot b = \begin{pmatrix} 0.61 \\ 0.99 \end{pmatrix}.$$

The resulting interpolation can be seen at the picture, where the given values b are interpolated with a green polygonal chain, while the red graph corresponds to the combination g. The computations were done in the system Maple using the command *leastsqrs(B,b)*. If you are enfriended with Maple (or some other similar software), try to do some experiments with similar tasks.

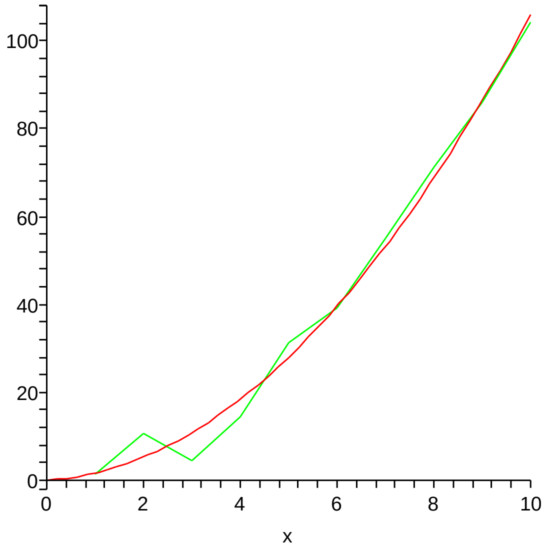
>From there we see that Au_k is an eigenvector of the matrix C with the eigenvalue λ_k. But that means that $Au_k = \lambda_k^A u_k$ for some number λ_k^A. Similarly we derive $Bu_k = \lambda_k^B u_k$ for some number λ_k^B. For the commutator of matrices A and B we then obtain

$$[A, B]u_k = ABu_k - BAu_k = \lambda_k^A \lambda_k^B u_k - \lambda_k^B \lambda_k^A u_k = 0$$

But that means that

$$[A, B]u = [A, B] \sum_k c_k u_k - \sum_k c_k [A, B] u_k = 0$$

and because u was arbitrary, that means that $[A, B] = 0$, which is a contradiction. □



3.45. Applications in quantum physics. In quantum physics we don't give to quantities any numerical value, as in classical physics, but a Hermitian operator. That is nothing but a Hermitian mapping, which can lead (and often does) to a linear transformation between unitary spaces of infinite dimension (we can imagine this as a matrix 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 it holds that $\hat{x}(v) = xv$, that is, it corresponds to the multiplication of the vector by the real number x. At the first glance this Hermitian operator is different from our cases of finite dimensions. Evidently every real number is an eigenvalue ($\hat{x}$ has the so-called continuous spectrum). Similarly, in place of speed (more precisely, momentum) we have the operator $\hat{p} = -i\frac{d}{dx}$. The eigenvectors are solution of the differential equation $-i\frac{dv}{dx} = \lambda v$. Even in this case is the spectrum continuous. That expresses the fact that the corresponding physical quantity is continuous (it 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 just have infinitely many eigenvalues.

3.46. Show that $\hat{x}$ and $\hat{p}$ are Hermitian and that

$$[\hat{x}, \hat{p}] = i$$

Solution. For any vector v it holds that

$$[\hat{x}, \hat{p}]v = \hat{x}\hat{p}v - \hat{p}\hat{x}v = x(-i\frac{dv}{dx}) + i\frac{d(xv)}{dx} - = iv$$

and from there we directly have our claim. □

3.47. Show that

$$[\hat{x} - \hat{p}, \hat{x} + \hat{p}] = 2i$$

Solution. Evidently we have that $[\hat{x}, \hat{x}] - 0$ and $[\hat{p}, \hat{p}] = 0$ and the rest follows from the linearity of the commutator from the previous exercise. □

3.48. Jordan form. Find the Jordan form of the matrix A. What is the geometric interpretation of this decomposition of the matrix?

i) $A = \begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix}$

ii) $A = \begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix}$

Solution. i)We 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 $\lambda_{1,2} = 1, 2$. Because 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, $-2x + y = 0$. That holds exactly for the multiples of the vector $(1, 2)$. Similarly we find out that 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 we then have $A = P \cdot J \cdot P^{-1}$. The inverse of P is $P^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$ and we obtain

$$\begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$$

This decomposition tells us that the matrix A determines such linear mapping that has in basis of the eigenvectors $(1, 2)$, $(1, 3)$ the aforementioned diagonal form. That means that in the direction $(1, 2)$ nothing is changing and in the direction $(1, 3)$ every vector is being stretched twice.

ii) Characteristic polynomial of the matrix A is in this case

$$|A - \lambda E| = \begin{vmatrix} -1-\lambda & 1 \\ -4 & 3-\lambda \end{vmatrix} = \lambda^2 - 2\lambda + 1 = 0$$

We obtain 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 has no 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 $A - E$, it is thus a solution of the system of equations

$$\left(\begin{array}{cc|c} -2 & 1 & 1 \\ -4 & 2 & 2 \end{array}\right) \sim \left(\begin{array}{cc|c} -2 & 1 & 1 \\ 0 & 0 & 0 \end{array}\right)$$

The solutions are the 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 and 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 $(1, 2)$. □

3.49. Find the Jordan form of the matrix A and write down the decomposition. What is 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 draw how the vectors $v = (3, 0)$, $A_1 v$ and $A_2 v$ 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 and both have in the basis of the vectors $(1, 2)$ and $(1, -1)$, that is,

$$\frac{1}{3}\begin{pmatrix} 5 & -1 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}^{-1}$$

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 vector $v = (3, 0)$ we obtain $v = (1, 2) + 2(1, -1)$ and for its images $A_1 v = (5, -2) = (1, 2) + 2 \cdot 2 \cdot (1, -1)$ and $A_2 v = (5, 4) = (2 + 1) \cdot (1, 2) + 2 \cdot (1, -1)$. □

F. Matrix decompositions

3.50. 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.

3.51. Find an LU-decomposition of the following matrix:

$$\begin{pmatrix} -2 & 1 & 0 \\ -4 & 4 & 2 \\ -6 & 1 & -1 \end{pmatrix}$$

Solution.

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

We first multiply the matrices that correspond to the Gaussian elimination, we thus obtain for the original matrix A, $XA = U$, where X is a lower triangular matrix given by the Gaussian reduction, U upper triangular. From this equality we have $A = X^{-1}U$, which is the desired decomposition (we thus have to compute the inverse of X). □

3.52. Find the LU-decomposition of the matrix $\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \\ - & 1 & -1 \end{pmatrix}$. ○

3.53. 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 approximation of the displayed objects by polyhedrons. These are bounded by planes and it is necessary to compute where exactly are the light rays reflected from these planes. From physics we know how are the rays reflected – the angle of impact equals the angle of reflection. With this topic we have already met in the exercise ||1.64||.

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. 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 picture we can directly derive that

$$v_R = v - 2\langle v, n\rangle n$$

and in our case we obtain $v_R = (-3, -2, -1)$. □

3.54. 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. We first compute $A^T A$:

$$A^T A = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}$$

and obtain a diagonal matrix. But we need to find such orthonormal basis under which the matrix is diagonal and the zero row is the last one. This can be clearly obtained by rotating through the right angle about the x-axis (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 (without much computation) found the decomposition $A^T A = VBV^T$, where B is diagonal with eigenvalues $(1, \frac{1}{4}, 0)$ on the diagonal. Because now we have $B = (AV)^T(AV)$, 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 $A = U\sqrt{B}V^T$

$$\begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

Geometrical interpretation of decomposition is the following: first, everything is rotated through the 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 $\frac{1}{2}$ and the result is the rotated through the right angle about the z-axis.

The polar decomposition $A = P \cdot W$ can be simply 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

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

and from that it follows that

$$\begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

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 we have

$$A^{(-1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

□

3.55. QR decomposition. QR decomposition of a matrix A is very useful in the case when we are given a system of linear equations $Ax = b$ which has no solution, but we need to find an approximation as good as possible. That is, we want to minimise $\|Ax - b\|$. According to the Pythagorean theorem we have $\|Ax-b\|^2 = \|Ax-b_{\|}\|^2 + \|b_{\perp}\|^2$, where b was decomposed into $b_{\|}$ that belongs to the range of the linear transformation A (that corresponds to the matrix A) and into $b_{\perp}$, that is perpendicular to this range. 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 through Gram-Schmidt orthonormalisation of the column of the matrix A. Then we have $Ax - b_{\|} = Q(Q^TAx - Q^Tb)$. The system in the parentheses has a solution, for which we obtain $\|Ax - b\| = \|b_{\perp}\|$, which is the minimal value. Furthermore, the matrix $R := Q^TA$ is upper triangular and therefore the approximate solution can be found very easily.

Find an approximate solution of the system

$$\begin{aligned} x + 2y &= 1 \\ 2x + 4y &= 4 \end{aligned}$$

Solution. We have a 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 thus orthonormalise the columns of A. We take the first of them and divide it by its size. 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 zero. Therefore we have $Q = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. The projector on the range of A is then $QQ^T = \frac{1}{5}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, next we compute

$$Q^Tb = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{9}{\sqrt{5}}$$

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 $Rx = Q^Tb$ and that in our case means $5x + 9y = 9$ (approximate solution is thus not unique). QR decomposition of the matrix A is then

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}\frac{1}{\sqrt{5}}\begin{pmatrix} 5 & 9 \end{pmatrix}$$

□

3.56. Minimise $\|Ax - b\|$ for $A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and write down the QR decomposition of the matrix A.

Solution. Normalised first column of the matrix A is 000 $e_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$. From the second column we subtract its component in the direction e_1. We have

$$\langle \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \rangle = -\frac{3}{\sqrt{6}}$$

and therefore we obtain

$$\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} - \langle \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \rangle \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$$

By this we have created an orthogonal vector, which we normalise and 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 (we can verify this by computing the determinant). 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 we compute

$$\begin{aligned} R = Q^T A &= \tfrac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \\ &= \tfrac{1}{\sqrt{6}} \begin{pmatrix} 6 & -3 & -3 \\ 0 & 3\sqrt{3} & -3\sqrt{3} \end{pmatrix} \end{aligned}$$

and

$$Q^T b = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

The solution of the equation $Rx = Q^T b$ is $x = y = z$. Multiples of the vector (1, 1, 1) thus minimise $\|Ax - b\|$.

The mapping given by the matrix A is a projection on the plane with a normal vector (1, 1, 1).

□

3.57. Linear regression. The knowledge we have obtained in this chapter can be successfully used practically for solving problems with linear regression. It is about finding the best approximation of some functional dependence using a linear function.

We are thus given a functional dependence in some points (for instance, we investigate the value of the property of people depending on their intelligence, the value of the property of their parents, number of mutual friends with Mr. Williams, ...), that is, $f(a_1^1, \dots, a_n^1) = y_1, \dots, f(a_1^k, a_2^k, \dots, a_n^k) = y_k$, $k > n$ (we have thus more equations than unknowns) and we want "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 $f(x_1, \dots, x_n) = b_1x_1 + b_2x_2 + \cdots + b_nx_n + c$. If we also define "best possible" by minimisation of

$$\sum_{i=1}^{k}\left(y_i - \sum_{j=1}^{n}(b_jx_j + c)\right)^2$$

with regard to the real constants $b_1, \dots, b_n, c$. Our goal is to find such linear combination of the columns of the matrix $A = (a_j^i)$ (with coefficients $b_1, \dots, b_n$), that has the smallest distance from the vector $(y_1, \dots, y_k)$ in $\mathbb{R}^k$, it is thus about finding an orthogonal projection of the vector $(y_1, \dots, y_k)$ on the subspace generated by the columns of the matrix A. Using the theorem 3.49 this projection is the vector $(b_1, \dots, b_n)^T = A^{(-1)}(y_1, \dots, b_n)$.

3.58. 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. Our system has no solution, since its matrix has rank 3, the extended matrix has rank 4. The best approximation of the vector $b = (1, 2, 0, -1)$ formed by the right sides of the equations can be thus obtained using the theorem 3.49 by the vector $A^{(-1)}b$. ($AA^{(-1)}b$ is then the best approximation – the perpendicular projection 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 $(A^TA)^{-1}A^T$. Thus we have

$$A^{(-1)} = \left(\begin{pmatrix} 2 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}\right)^{-1}\begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \left(\begin{pmatrix} 10 & 5 & 10 \\ 5 & 3 & 6 \\ 10 & 6 & 15 \end{pmatrix}\right)^{-1}\begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 3/5 & -1 & 0 \\ -1 & 10/3 & -2/3 \\ 0 & -2/3 & 1/3 \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1/5 & -2/5 & 1/5 & 3/5 \\ 0 & 1/3 & 2/3 & -5/3 \\ 0 & 1/3 & -1/3 & 1/3 \end{pmatrix}$$

The desired $\mathbf{x}$ equals

$$A^{(-1)}b = (-6/5, 7/3, 1/3)^T.$$

The projection (the best possible approximation of the column of the right sides) is then the vector $(3/5, 32/15, 4/15, -13/15)$. □

G. Additional exercises for the whole chapter

3.59. Model of evolution of a whale population. For evolution of a population females are important, and for them 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 infertile; young fertile females; adult females with highest fertility and postclimacterial females which are not fertile anymore (but are still important with respect to taking care of newborns and food gathering).

We model the evolution of such population in time. For a time unit we choose time it takes to reach adulthood. Newborn female which survives this interval becomes fertile. The evolution of a young female to full fertility and to postclimacterial state depends on the environment. That is, transition to next category is a random event. Analogously, death of an individual is also a random event. Young fertile female has per unit interval less children than adult female. Let us 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 on a unit area (the so-called population density), or as a total biomass and similarly. Further denote 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 adult female becomes old. Another random event is dying (positively formulated: survival) of females that do not move to the next category – we denote the probabilities respectively 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 as a probability from the interval [0, 1]. Young female can survive, reach adulthood or die; these events are mutually exclusive, together they form a sure event and cannot be excluded. Thus we have $p_2 + q_2 < 1$. From similar reasons we have $p_3 + q_3 < 1$. Finally, we denote by f_2 and f_3 the average number of daughters of a young and adult female, respectively. These parameters satisfy $0 < f_2 < f_3$.

Expected number of newborn females in the next time interval is the sum of daughters of young and of adult females, that is

$$x_1(t+1) = f_2 x_2(t) + f_3 x_3(t).$$

We denote for a while by $x_{2,1}(t+1)$ the amount of young females in time $t+1$, which were in the previous time interval, that is, in time t, juvenile, and by $x_{2,2}(t+1)$ the amount of young females, that were already in time t fertile, 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 $x_{2,1}(t+1)/x_1(t)$, and similarly we can express the probability q_2 as the ratio $x_{2,2}(t+1)/x_2(t)$. Because young females in time $t+1$ are exactly those that survived the juvenile stage and those that already were fertile, did survive and did not evolve, it holds that

$$x_2(t+1) = x_{2,1}(t+1) + x_{2,2}(t+1) = p_1 x_1(t) + q_2 x_2(t).$$

Analogically we derive the expected number of fully fertile females

$$x_3(t+1) = p_2 x_2(t) + q_3 x_3(t)$$

and the expected number of postclimacterial females by

$$x_4(t+1) = p_3 x_3(t) + q_4 x_4(t).$$

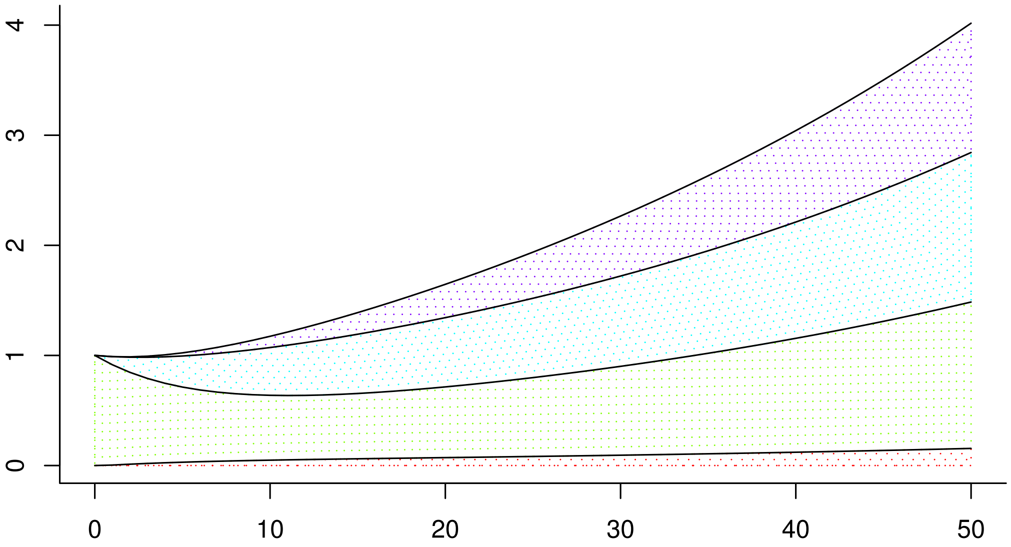


FIGURE 1. Evolution of a population of orca whale. On the horizontal axis the time is in years, on the vertical axis is the size of the population. Individual areas depict the number of juvenile, young, adult and old females respectively, from bellow.

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 the matrix form

$$x(t+1) = Ax(t).$$

Using this matrix difference equation we can easily compute the expected number of whale females in individual categories, if we know the distribution of population at some initial time.

Specifically, for the population of orca whales the following parameters were observed:

$$\begin{aligned} &p_1 = 0{,}9775, \quad q_2 = 0{,}9111, \quad f_2 = 0{,}0043, \\ &p_2 = 0{,}0736, \quad q_3 = 0{,}9534. \quad f_3 = 0{,}1132, \\ &p_3 = 0{,}0452, \quad q_4 = 0{,}9804; \end{aligned}$$

Time interval is in this case one year.

If we start at the time $t = 0$ with unit measure of young female in some unoccupied area, that is, with the vector $x(0) = (0, 1, 0, 0)^T$, we can compute

$$x(1) = \begin{pmatrix} 0 & 0{,}0043 & 0{,}1132 & 0 \\ 0{,}9775 & 0{,}9111 & 0 & 0 \\ 0 & 0{,}0736 & 0{,}9534 & 0 \\ 0 & 0 & 0{,}0452 & 0{,}9804 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}0043 \\ 0{,}9111 \\ 0{,}0736 \\ 0 \end{pmatrix},$$

$$x(2) = \begin{pmatrix} 0 & 0{,}0043 & 0{,}1132 & 0 \\ 0{,}9775 & 0{,}9111 & 0 & 0 \\ 0 & 0{,}0736 & 0{,}9534 & 0 \\ 0 & 0 & 0{,}0452 & 0{,}9804 \end{pmatrix} \begin{pmatrix} 0{,}0043 \\ 0{,}9111 \\ 0{,}0736 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}01224925 \\ 0{,}83430646 \\ 0{,}13722720 \\ 0{,}00332672 \end{pmatrix}$$

and we can carry on. The results of the computation can be also expressed graphically; see the picture ‖1‖. Try by yourself a computation and graphical depiction of the results even 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 gradually on constant values.

The matrix A thus has the eigenvalues

$$\lambda_1 = 1{,}025441326,\ \lambda_2 = 0{,}980400000,\ \lambda_3 = 0{,}834222976,\ \lambda_4 = 0{,}004835698,$$

eigenvector associated with the largest eigenvalue λ_1 is

$$w = (0{,}03697187,\ 0{,}31607121,\ 0{,}32290968,\ 0{,}32404724);$$

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 in individual categories in the population after certain time of evolution, and compare it with the components of the eigenvector w. They will appear very close, this is caused by the fact that A has only single eigenvalue that has the greatest absolute value and by the fact that the vector space generated by the eigenvectors associated with the eigenvalues λ_2, λ_3, λ_4 has with the non-negative orthant intersection only the zero vector. The structure of the matrix A itself does not ensure such easily predictable evolution, because it is a so-called reducible matrix (see ??).

3.60. Model of growth of population of teasels *Dipsacus sylvestris*. This plant can be seen in four stages. Either as a blossoming plant or as rosette of leaves, while with the rosette there are three sizes – small, medium and large. The life cycle of this monoicous perennial plant can be described as follows.

Blossoming plant produces in late summer some number of seeds and dies. From the seeds, some sprout already 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, small or medium rosette can during the year grow, and any rosette can stay in its category or die (wither, be eaten by insects, etc.) Medium or large rosette can in the next year burst into a flower. 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. The 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.

Let us 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 could lead us to 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 in the same year changed either into rosettes or they spend winter, thus they do not form an individual category. Let us thus 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 the seeds stays dormant for the first year equals the probability that the seed does not sprout into any 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 have been dormant for one year stays dormant for the second year equals the probability that the dormant seed does not sprout into any 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

$$x_2(t+1) = 0{,}966x_1(t).$$

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

$$\begin{aligned} x_4(t+1) =& 0{,}007x_1(t) + 0,125x_3(t) + 0{,}238x_4(t) + 0{,}070 \cdot 431x_6(t) = \\ =& 0{,}007x_1(t) + 0,125x_3(t) + 0{,}238x_4(t) + 30,170x_6, \end{aligned}$$

$$\begin{aligned} x_5(t+1) =& 0{,}245x_4(t) + 0,167x_5(t) + 0{,}002 \cdot 431x_6(t) = \\ =& 0{,}245x_4(t) + 0,167x_5(t) + 0{,}862x_6(t). \end{aligned}$$

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 the 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 $t = 0$, 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, $n(t) = \sum_{i=1}^{6} x_i(t)$, 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 that we have put into some locality one blossoming flower, are given in the table ‖1‖. Unlike the whale population, the image would not be very clear, as 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 the eigenvalues

$$\begin{array}{ll} \lambda_1 = 2{,}3339 & \lambda_4 = 0,1187 + 0,1953\mathrm{i} \\ \lambda_2 = -0{,}9569 + 1{,}4942\mathrm{i} & \lambda_5 = 0,1187 - 0,1953\mathrm{i} \\ \lambda_3 = -0{,}9569 - 1{,}4942\mathrm{i} & \lambda_6 = -0,1274 \end{array}$$

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 equal to one. We see that with increasing time t the relative increment in the size of population approaches the eigenvalue λ_1, 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 at the

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	$\frac{x_1(t)}{n(t)}$	$\frac{x_2(t)}{n(t)}$	$\frac{x_3(t)}{n(t)}$	$\frac{x_4(t)}{n(t)}$	$\frac{x_5(t)}{n(t)}$	$\frac{x_6(t)}{n(t)}$	$\frac{n(t+1)}{n(t)}$
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.

same positions as A is primitive. The evolution of the population thus necessarily approaches a stable structure.

3.61. Nonlinear model of population. Investigate in detail the evolution of the population for a non-linear model from the text book (1.12) and the values and $K = 1$ 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 first members 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

We see that 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

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:

□

3.62. In a lab an experiment is being carried on with the same probability of success and failure. If the experiment succeeds, the probability of the success of the second experiment is $0,7$. If the first experiment fails, the probability of the success of the second experiment is only $0,6$.

This process goes on, that is, if the previous experiment was successful, the probability of the next success is $0,7$ and if the previous experiment was a failure, then the probability of the next success is $0,6$. For any $n \in \mathbb{N}$ determine the probability that the n-th experiment is successful.

Solution. Let us 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 it is

$$x_1 = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

and clearly 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

$$x_{n+1} = T \cdot x_n, \quad n \in \mathbb{N}, \tag{3.7}$$

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 (‖3.7‖) we directly have

$$x_{n+1} = T \cdot T \cdot x_{n-1} = \cdots = T^n \cdot x_1, \quad n \geq 2,\ n \in \mathbb{N}. \tag{3.8}$$

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 follows for instance from the fact that the trace (the sum of the elements on the diagonal) equals to 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 (‖3.7‖) and from (‖3.8‖) 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 $2/3$. □

3.63. Student on a student dormitories 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 that he invites on the party-in-progress his friend which lives at the end of the hall. But, at the other end of the hall there lives somebody he definitely does not want to invite. But he is so „tired“, that he realises 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. ○

3.64. Let $n \in \mathbb{N}$ of persons be playing the so-called "silent post". For simplicity assume that the first person whispers to the second person exactly one (arbitrarily chose) of the words „yes“, „no“. The second person then whispers to the third one that of the words „yes“, „no“ the second person thinks that the first person whispered. This then continues to the n-th person. If the probability that the word changes (on purpose, accidentally) to the other word during one transmission equals $p \in (0, 1)$, determine for big $n \in \mathbb{N}$ the probability that the n-th person correctly receives the word transmitted by the first person.

Solution. We can view this problem as a Markov chain with two states called Yes and No, and we say that the process is in the Yes state in the 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 don’t 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 we know that for big $n \in \mathbb{N}$ 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 big number of participants thus holds that roughly half of them hears „yes“ (we repeat that this does not depend on the initial word).

For completeness let us determine what would be the result if we assumed that the probability of change from „yes“ to „to“ is for any person equal to $p \in (0, 1)$ and the probability of change from „no“ to „yes“ is equal to (in general distinct) $q \in (0, 1)$. In this case for the same order of the states we obtain a probabilistic matrix

$$T = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix},$$

which leads to (for big $n \in \mathbb{N}$) 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 holds that it does not depend on the initial state, because the people decide about what the transmitted information is; more precisely, the people themselves decide about the frequency of appearance of „yes“ and „no“, if there is enough of them (and there is no checking present).

Let us further add that the obtained result was experimentally confirmed. In psychological experiment there was an individual repeatedly exposed to an event that could have been interpreted in two ways, and it was being done in time intervals that ensured that the subject still remembered the previous event. See for instance „T. Havr’anek et al.: *Matematika pro biologick’e a l’ekařsk’e 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 $p, q \in (0, 1)$. □

3.65. 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. But you await a big number of games (and for each of them you can choose a different opponent). And of course you want reach the winning ratio as big as possible. Consider these two strategies:

1. For the first game choose the opponent randomly. If you win some game, carry on with the same opponent; if you lose the game, change the opponent.
2. For the first two games, choose (one) opponent randomly. Then for the next two games, if you lost both the previous games, change the opponent, otherwise stay with the same.

Which of the strategies is better?

Solution. Both strategies are 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 elements positive, and thus it suffices to find the probabilistic vector x_∞, which is associated with the eigenvalue 1. It holds that

$$x_\infty = \left(\frac{3}{5}, \frac{2}{5}\right)^T.$$

Its components correspond to the probabilities that after a long row 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 opponents. Because

$$\frac{2}{5} = \frac{3}{5} \cdot \frac{1}{2} + \frac{2}{5} \cdot \frac{1}{4},$$

there will be roughly 40 %.

For the second strategy, let us use the states „two games in a row with A" and „two games in a row with B" that lead to the probabilistic transition matrix

$$\begin{pmatrix} 3/4 & 9/16 \\ 1/4 & 7/16 \end{pmatrix}.$$

We easily determine that now it is

$$x_\infty = \left(\frac{9}{13}, \frac{4}{13}\right)^T.$$

Against the worse opponent we would then play $(9/4)$-times more frequently than against the better one. Let us recall that for the first strategy it was $(3/2)$-times more frequently. The second strategy is thus better. Let us also note that for the second strategy roughly 42,3 % of the games are winning – it suffices to enumerate

$$0,423 \doteq \frac{11}{26} = \frac{9}{13} \cdot \frac{1}{2} + \frac{4}{13} \cdot \frac{1}{4}.$$

□

3.66. Petr regularly meets his friend. But he is „well-known“ for his bad timekeeping. Bud he is trying to change, thus it holds that in half of the cases he comes on time and in one tenth of the cases he comes even sooner than he should, if he was late for the last meeting. But if he was on time or sooner for the last meeting, he returns back to his „carelessness“ and with probability $0,8$ comes late, and with only $0,2$ he is on time. What is the probability that on the 20-th meeting he comes late, when on the eleventh he was on time?

Solution. Clearly it is a Markov process with states „Petr came late“, „Petr came on time“, „Petr came 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 $(0, 1, 0)^T$ (we surely know that Petr came 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). Let us 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 there we see that it really does not depend on whether he came on the eleventh meeting late (first column), on time (second) or sooner (third). □

3.67. 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 the half. The result of the previous game partially determines the next game. If a week ago the student A has won, then with the probability $3/4$ wins again and with probability $1/4$ it is a draw. Draw is repeated with the probability $2/3$ and with probability $1/3$ the next game is won by B. If the student B won a game, then with the probability $1/2$ he wins again and with probability $1/4$ student A is the winner of 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. We are actually given a Markov process with the states „the student A wins“, „the game ends with a draw, „the 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 big number $n \in \mathbb{N}$ of steps (weeks). The matrix T is primitive, because

$$T^2 = \begin{pmatrix} 9/16 & 1/12 & 5/16 \\ 17/48 & 19/36 & 17/48 \\ 1/12 & 7/18 & 1/3 \end{pmatrix}.$$

It thus suffices to find the probabilistic eigenvector x_∞ of the matrix T associated with the eigenvalue 1. It is easy to compute that

$$x_\infty = \left(\frac{2}{7}, \frac{3}{7}, \frac{2}{7}\right)^T .$$

We know that the vector x_∞ differs only very slightly from the probabilistic vector for big n and also does not depend on the initial state, that is, for big $n \in \mathbb{N}$ we can set

$$T^n \approx \begin{pmatrix} 2/7 & 2/7 & 2/7 \\ 3/7 & 3/7 & 3/7 \\ 2/7 & 2/7 & 2/7 \end{pmatrix} .$$

The desired probability is the element of this matrix on the second position in the first column (the second component of the vector x_∞). Thus we have (quite quickly) found the result $3/7$. □

Solutions of the exercises

3.2. Daily diet should contain 3, 9 kg of hay and 4, 3 kg of oat. The costs per foal are then 13, 82 Kč.

3.3.

3.12.

$$x_n = \frac{1}{\sqrt{21}}\left(\frac{3+\sqrt{21}}{2}\right)^n - \frac{1}{\sqrt{21}}\left(\frac{3-\sqrt{21}}{2}\right)^n.$$

3.13. $x_n = 2\sqrt{3}\sin(n \cdot (\pi/6)) - 4\cos(n \cdot (\pi/6))$.

3.14. $x_n = -3(-1)^n - 2\cos(n \cdot (2\pi/3)) - 2\sqrt{3}\sin(n \cdot ((2\pi/3))$.

3.15. $x_n = (-1)^n(-2n^2 + 8n - 7)$.

3.24. 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, $a = 1/4$.

3.27.

$$\begin{pmatrix} \frac{5}{6} & \frac{1}{5} \\ \frac{1}{6} & \frac{4}{5} \end{pmatrix}.$$

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.

3.30. As in (‖3.29‖) the game ends after three bets. Thus all the powers of A, starting with A^3, are identical.

$$A^{100} = A^3 = \begin{pmatrix} 1 & 7/8 & 3/4 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/8 & 1/4 & 1/2 & 1 \end{pmatrix}$$

3.40. We can use the result of the exercise called Ruining of the player. The probability that the first department is cancelled is according to this 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 plug in $p = 1 - 0.54$, $y = 10/2$ and $x = 40/2$ to (‖3.6‖). It is thus more clever to choose the smaller department.

3.50.

- The claim holds. ($B := A^TA$, 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 does not hold. Consider for instance $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

3.52.

$$\begin{pmatrix} 1 & 0 & 0 & \\ & 1 & -2 & 0 \\ 0 & 1 & 1 & \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

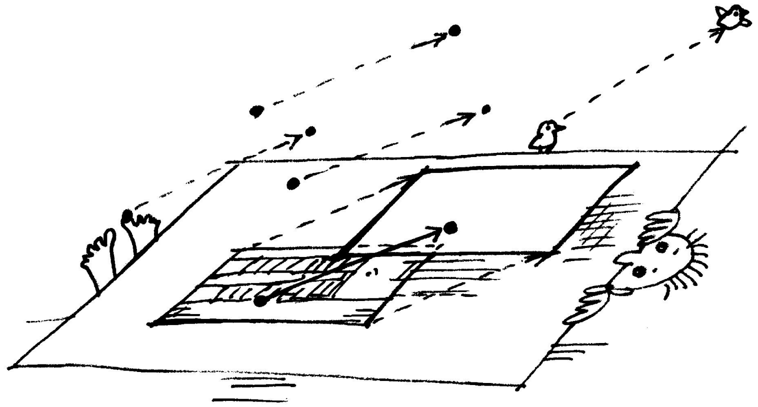
3.63. Again it is a special case of the Ruining of the player. It suffices to reformulate the statement accordingly. For $p = 0, 47$, $y = 20$ and $x = 20$ from (‖3.6‖) follows the result

$$0,917 \doteq \frac{1 - \left(\frac{0{,}47}{1-0{,}47}\right)^{20}}{1 - \left(\frac{0{,}47}{1-0{,}47}\right)^{40}}.$$

CHAPTER 4

Analytic geometry

position, incidence, projection
– and we return to matrices again...



A. Affine geometry

4.1. Find the parametric equation for a line in $\mathbb{R}^3$ given by equations

$$\begin{array}{rcrcrcl} x & - & 2y & + & z & = & 2, \\ 2x & + & y & - & z & = & 5. \end{array}$$

Solution. It is obviously sufficient to solve the equation system. However we can use different approach. We need to find non-zero direction vector ortogonal to normal vectors $(1, -2, 1)$, $(2, 1, -1)$. Cross product

$$(1, -2, 1) \times (2, 1, -1) = (1, 3, 5)$$

gives us such vector. We can notice that triple

$$(x, y, z) = (2, -1, -2)$$

satisfies the respective system and we obtain the solution

$$[2, -1, -2] + t\,(1, 3, 5)\,, \quad t \in \mathbb{R}.$$

□

4.2. 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.

Now we come back to our 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. 1.23. First we will be interested in properties of objects in the Euclidean space, delimited by points, straight lines, planes etc. The essential point will be to clarify how their properties are related to the notion of vectors and whether they depend on the notion of length of vectors.

In the next part, we will use linear algebra for the study of objects which are defined in a nonlinear way. To do this we will need a little bit more from the theory of matrices again. The results will be important later on, while discussing the technique for optimalization, i.e. searching for extrema of functions.

At the end of this chapter we show how the projectivization of affine spaces help us to get a simplification and stability of algorithms typical for computer graphics.

1. Affine and euclidean geometry

While we were clarifying the structure of solutions of linear equations in the first part of the previous chapter we found out in paragraph 3.1 that all solutions of nonhomogeneous systems of linear equations does not form vector spaces but always arise in such a way that to an one particular solution we add the vector space of solutions of 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, as we have seen in paragraph 3.14 already.

4.1. Affine spaces. A direction how to deal with the theory is given already in the discussion about the geometry of the plane, c.f. paragraph 1.25 and further. There we described straight lines and points as sets of solutions of systems of linear equations. Any line was considered as a one-dimensional subspace, although its points were described by two coordinates. Parametrically, the line was defined by the sum of a single point (i.e. to a pair of coordinates) and multiples of a fixed direction vector. Now we will proceed in the same way in arbitrary dimension.

Standard affine space

Standard affine space $\mathcal{A}_n$ is a set of all points in $\mathbb{R}^n = \mathcal{A}_n$ together with an operation which to a point $A = (a_1, \dots, a_n) \in \mathcal{A}_n$ and a vector $v = (v_1, \dots, v_n) \in \mathbb{R}^n = V$ assigns the point

$$A + v = (a_1 + v_1, \dots, a_n + v_n) \in \mathbb{R}^n = \mathcal{A}_n.$$

Solution. Our task is to find a system of equations with 4 variables x, y, z, u (because dimension of the space is 4) which are satisfied by the coordinates of precisely those points which lie in the plane. Note that sought system must contain $2 = 4 - 2$ linearly independent equations. We solve the problem by so called elimination of parameters. 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 $t, s \in \mathbb{R}$. We can express the system as matrix

$$\left(\begin{array}{cc|cccc|c} 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 3 \\ 1 & -2 & 0 & 0 & -1 & 0 & 2 \\ 0 & 2 & 0 & 0 & 0 & -1 & 5 \end{array}\right),$$

where the first two columns are direction vectors of the plane, followed by negative identity matrix and finally the last column is vector of coordinates of point $[0, 3, 2, 5]$. We expressed the system in such a way so that it is a system in t, s, x, y, z, u and we move all the unknown variables to the one side of the equations. We transform obtained matrix using elementary row operations in order to get as much zero-rows on the left-hand side of the first vertical line. 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 we obtain

$$\left(\begin{array}{cc|cccc|c} 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 3 \\ 1 & -2 & 0 & 0 & -1 & 0 & 2 \\ 0 & 2 & 0 & 0 & 0 & -1 & 5 \end{array}\right) \sim$$
$$\sim \left(\begin{array}{cc|cccc|c} 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 4 & -1 & 0 & -10 \\ 0 & 0 & 0 & -2 & 0 & -1 & 11 \end{array}\right).$$

Which implies result

$$\begin{array}{rcrcrcrcrcl} x & + & 4y & - & z & & & - & 10 & = & 0, \\ & & -2y & & & - & u & + & 11 & = & 0. \end{array}$$

Coefficients on the right-hand side of the first vertical line, respective to the rows which are zero-rows on the left-hand side of that line, are the coefficients of general equations of a planes.

Note that if we expressed the original system as a matrix

$$\left(\begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 0 & 2 & 5 \end{array}\right),$$

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 $\mathcal{A}_n$.

We notice a danger of several formal ambiguities. We are using the same symbol "+" for two different operations: for adding a vector from the difference space to a point in the affine space, and for for summing vectors in the difference space $V = \mathbb{R}^n$. Also we do not introduce specific letters for the set of points in the affine space, i.e. $\mathcal{A}_n$ denotes both this set of points and also the whole structure defining the affine space.

Why do we actually want to 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 going on fundamental formal step to understanding the geometry in $\mathbb{R}^n$: The thing is that the geometric objects like straight lines, points, planes etc. do not depend directly on the vector space structure of the set $\mathbb{R}^n$, and do not depend at all on the fact that we are working with n–tuples of scalars. We only need to know what it means to move "straight in a given direction". For instance, we consider the affine plane as an unbounded board without chosen coordinates but with the possibility to move about a given vector. When we switch to such abstract view, we will be able to discuss the "plane geometry" for two-dimensional subspaces, i.e. planes in higher-dimensional spaces, the geometry of "Euclidean space" for three-dimensional subspaces etc., without the need to work with k–tuples of coordinates.

This point of view is present in the following definition:

4.2. Definition. The affine space $\mathcal{A}$ with the difference space V is a set of *points* $\mathcal{P}$, together with the map

$$\mathcal{P} \times V \to \mathcal{P}, \quad (A, v) \mapsto A + v,$$

where V is a vector space and our map satisfies the properties (1)–(3) from the definition of the standard affine space.

So for a fixed vector $v \in V$ we get a *translation* $\tau_v : \mathcal{A} \to \mathcal{A}$ as the restricted map

$$\tau_v : \mathcal{P} \simeq \mathcal{P} \times \{v\} \to \mathcal{P}, \quad A \mapsto A + v.$$

By the *dimension* of an affine space $\mathcal{A}$, we mean the dimension of its difference space.

In sequel we do not distinguish accurately between denoting the set of points $\mathcal{A}$ and the set of vectors $\mathcal{P}$, we talk about points and vectors of the affine space $\mathcal{A}$ instead.

It follows immediately form the axioms that for arbitrary points A, B, C in the affine space $\mathcal{A}$

$$A - A = 0 \in V \tag{4.1}$$
$$B - A = -(A - B) \tag{4.2}$$
$$(C - B) + (B - A) = C - A. \tag{4.3}$$

Indeed, (4.1) follows from the fact that $A+0 = 0$ and that such vector is unique (the first and the third defining property). By adding where x, y, z, u remains on the left-hand side of the equations, similar transformation

$$\left(\begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 0 & 2 & 5 \end{array}\right) \sim \left(\begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 3 \\ -1 & -4 & 1 & 0 & 0 & 0 & -10 \\ 0 & 2 & 0 & 1 & 0 & 0 & 11 \end{array}\right)$$

gives us the result

$$\begin{array}{rcrcrcrcr} -x & - & 4y & + & z & & & = & -10, \\ & & 2y & & & + & u & = & 11. \end{array}$$

When expressing system as a matrix, it is important to take into consideration whether the vertical line separates left-hand side from right-hand side. As we saw in this exercise, parameter elimination method can be long-winded and it is not difficult to make a mistake along the way.

Another solution All we wanted to obtain in fact, are two linearly independent normal vectors, i.e. vectors perpendicular to $(1, 0, 1, 0)$, $(2, -1, -2, 2)$. If we "guessed" that these vectors could be for example $(0, 2, 0, 1)$, $(-1, 0, 1, 2)$, inputting $x = 0$, $y = 3$, $z = 2$, $u = 5$ to the equations

$$\begin{array}{rcrcrcrcr} & & 2y & & & + & u & = & a, \\ -x & & & + & z & + & 2u & = & b \end{array}$$

we get $a = 11$, $b = 12$, and the sought implicit expression is

$$\begin{array}{rcrcrcrcr} & & 2y & & & + & u & = & 11, \\ -x & & & + & z & + & 2u & = & 12. \end{array}$$

□

4.3. Find a parametric equation of the plane passing through points

$$A = [2, 1, 1], \quad B = [3, 4, 5], \quad C = [4, -2, 3].$$

Then find a parametric equation of the open half-plane containing the point C and bounded by line going through the points A, B.

Solution. We need one point and two (linearly independent) vectors lying in this plane for the parametric equation of the plane. It is enough to choose the point A and vectors $B - A = (1, 3, 4)$ and $C - A = (2, -3, 2)$, which are obviously independent. A point $[x, y, z]$ lies in the plain if and only if there exist numbers $t, s \in \mathbb{R}$ so that

$$x = 2 + 1 \cdot t + 2 \cdot s, \quad y = 1 + 3 \cdot t - 3 \cdot s, \quad z = 1 + 4 \cdot t + 2 \cdot s;$$

which means the parametric equation is

$$[2, 1, 1] + t\,(1, 3, 4) + s\,(2, -3, 2)\,, \quad t, s \in \mathbb{R}.$$

Setting $s = 0$ gives us a line passing through points A, B. For $t = 0$, $s \geq 0$ we get a ray with initial point A and passing through C. Particular but arbitrarily choosen $t \in \mathbb{R}$ and variable $s \geq 0$ gives us a ray initiated on the border line and going through the half-plane in

successively $B - A$ and $A - B$ to A, according to the second defining property we obtain obviously A again. So we added the null vector which proves (4.2). Similarly, (4.3) follows from the defining property 4.1 (2) and the uniqueness.

Let us remark 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 we get for every point $A \in \mathcal{A}$ a unique expression

$$A = A_0 + x_1 u_1 + \cdots + x_n u_n.$$

We talk about an *affine coordinate system* $(A_0; u_1, \dots, u_n)$ given by the *origin of the affine coordinate system* A_0 and the basis $\underline{u}$ of the corresponding difference space, or also about an *affine frame* $(A_0, \underline{u})$.

We can summarize the situation as follows: 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 $\mathcal{A}_n$.

4.3. Affine subspaces. If we choose only such points in $\mathcal{A}$ which have some of in advance chosen coordinates equal to zero (for instance the last one), we obtain again a set which behaves as an affine space. Indeed, this is the spirit of the following definition of the so called 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 for any $A \in \mathcal{Q}$, $v \in W$ we have $A + v \in \mathcal{Q}$.

It is important to include both of the conditions in the definition since it is easy to find examples of sets which satisfy the first condition but not the second one. Have a think about a straight line in the plane with one removed point.

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.

In particular, $V = Z(\mathcal{A})$ and every affine subspace $\mathcal{Q} \subset \mathcal{A}$ itself satisfies the axioms for an affine space with the difference space $Z(\mathcal{Q})$.

Directly from the definitions we also get that the intersection of any set of affine subspaces is either an affine subspace or the empty set.

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 the subset M.

Affine hull and parametric description of a subspace

The affine subspaces can be nicely described by their difference spaces if we choose a point $A_0 \in M$ in a generating set M. Indeed, we get $\langle M \rangle = \{A_0 + v;\, v \in Z(M) \subset Z(\mathcal{A})\}$, i.e. to generate the affine subspace we take the vector subspace $Z(M)$ in the difference space generated by all differences of points in M, and we add this vector space to an arbitrary point in M. We talk also about the *affine hull* of the set of points M in $\mathcal{A}$.

which point C lies. That means that the sought open half-plane can be expressed parametrically as

$$[2,1,1]+t\,(1,3,4)+s\,(2,-3,2)\,,\qquad t\in\mathbb{R},\ s>0.$$

□

4.4. Determine relative position of lines

$$p:[1,0,3]+t\,(2,-1,-3)\,,\qquad t\in\mathbb{R},$$
$$q:[1,1,3]+s\ \,(1,-1,-2)\,,\qquad s\ \in\mathbb{R}.$$

Solution. We will find common points of given lines (subspaces intersection). We get a system

$$\begin{array}{rcrcrcr} 1 & + & 2t & = & 1 & + & s,\\ 0 & - & t & = & 1 & - & s,\\ 3 & - & 3t & = & 3 & - & 2s. \end{array}$$

>From the first two equations we get that $t=1$, $s=2$. However, this does not satisfy the third equation. The system does not have a solution. Direction vector $(2,-1,-3)$ of the line p is not a multiple of direction vector $(1,-1,-2)$ of the line q which means that the lines are not parallel. Hence, they are skew lines. □

4.5. Find all numbers $a\in\mathbb{R}$ so that lines

$$p:[4,-4,8]+t\,(2,1,-4)\,,\qquad t\in\mathbb{R},$$
$$q:[a,6,-5]+s\ \,(1,-3,3)\,,\qquad s\ \in\mathbb{R}$$

are intersecting.

Solution. Lines are intersecting if and only if the system

$$\begin{array}{rcrcrcr} 4 & + & 2t & = & a & + & s,\\ -4 & + & t & = & 6 & - & 3s,\\ 8 & - & 4t & = & -5 & + & 3s \end{array}$$

has exactly one solution. Expressing the system as a matrix (the first column corresponding to t, the second to s), we solve

$$\left(\begin{array}{cc|c} 2 & -1 & a-4\\ 1 & 3 & 10\\ -4 & -3 & -13 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 3 & 10\\ 2 & -1 & a-4\\ -4 & -3 & -13 \end{array}\right)$$
$$\sim \left(\begin{array}{cc|c} 1 & 3 & 10\\ 0 & -7 & a-24\\ 0 & 1 & 3 \end{array}\right).$$

We see that the system has exactly one solution if and only if the second row is a multiple of the third row. This property is satisfied only for $a=3$. Let us add that the point of intersection of the lines is $[6,-3,4]$.

□

On the other hand, whenever we choose a subspace U in the difference space $Z(\mathcal{A})$ and a fixed point $A\in\mathcal{A}$ the subset $A+U$, created by all possible sums of the point A and all vectors in U, is an affine subspace. This approach leads to the notion of parametrization of subspaces:

Let $\mathcal{Q}=A+Z(\mathcal{Q})$ is an affine subspace in $\mathcal{A}_n$ and $(u_1,\dots,u_k)$ is a basis of $Z(\mathcal{Q})\subset\mathbb{R}^n$. Then the expression of the subspace

$$\mathcal{Q}=\{A+t_1u_1+\dots+t_ku_k;\,t_1,\dots,t_k\in\mathbb{R}\}$$

is called the *parametric description* of the subspace $\mathcal{Q}$.

We have seen already another way how to prescribe 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 our subspace $\mathcal{Q}$ into the system of equations we get the right-hand side of the nonhomogeneous system with the same matrix, and the whole 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 general proposition says that we can prescribe all affine subspaces in this way, and so it also shows the geometric nature of solutions of systems of linear equations.

4.4. Theorem. *Let $(A_0;\underline{u})$ be an affine coordinate system in a 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. Let us consider an arbitrary solvable system of $n-k$ linearly independent equations $\alpha_i(x)=b_i$, where $b_i\in\mathbb{R}$, $i=1,\dots,n-k$. If $A=(a_1,\dots,a_n)^T\in\mathbb{R}^n$ is a fixed solution of this (nonhomogeneous) system and if $U\subset\mathbb{R}^n$ is the vector space of all solutions of the homogenized system $\alpha_i(x)=0$, then the dimension of U is k and the subset of all solutions of the given system is of the form $\{B;\,B=A+(y_1,\dots,y_n)^T,\,y=(y_1\dots,y_n)^T\in U\}\subset\mathbb{R}^n$, c.f. 3.1. So the corresponding affine subspace is described parametrically by the initial coordinates $(A_0;\underline{u})$.

In the opposite direction, let us consider an arbitrary affine subspace $\mathcal{Q}\subset\mathcal{A}_n$, let us choose a point B therein, and let us consider this point to be the origin of an affine coordinate system $(B,\underline{v})$ for the affine space $\mathcal{A}$. Since $\mathcal{Q}=B+Z(\mathcal{Q})$, we need to describe the difference space of the subspace $\mathcal{Q}$ as a subspace of solutions of a homogeneous system of linear equations. Therefore let us choose a basis $\underline{v}$ of $Z(\mathcal{A})$ such that the first k vectors form a basis of $Z(\mathcal{Q})$. Then in these coordinates the vectors $v\in Z(\mathcal{Q})$ are given by equations

$$\alpha_j(v)=0,\qquad j=k+1,\dots,n,$$

where α_i are linear forms from the so called dual basis to $\underline{v}$, i.e. functions which assign to a vector the corresponding coordinates in our basis $\underline{v}$.

Hence our vector subspace $Z(\mathcal{Q})$ of dimension k in the n-dimensional space $\mathbb{R}^n$ is given indeed as a solution of a homogeneous system of $n-k$ independent equations. The description of the chosen affine subspace in our newly chosen coordinate system $(B;\underline{v})$ is therefore given by a system of homogeneous linear equations.

4.6. In $\mathbb{R}^3$, determine the relative position of a line p defined implicitly by

$$\begin{array}{rcrcrcr} x & + & y & - & z & = & 4, \\ x & - & 2y & + & z & = & -3 \end{array}$$

and a plane $\varrho : y = 2x - 1$.

Solution. Normal vector ϱ is $(2, -1, 0)$ (consider $\varrho : 2x - y + 0z = 1$). It can be seen that

$$(1, 1, -1) + (1, -2, 1) = (2, -1, 0),$$

which means that the normal vector of the plane ϱ is a linear combination of the p normal vectors. Vector defining the line (given by non-zero direction vector perpendicular to the normal vectors) lies in a subspace of the plane ϱ (direction vector is perpendicular to the vector $(2, -1, 0)$). Therefore we know that the line p is parallel to the plane ϱ. Now we have to find whether they intersect (meaning that p lies in ϱ). System of equations

$$\begin{array}{rcrcrcr} x & + & y & - & z & = & 4, \\ x & - & 2y & + & z & = & -3, \\ 2x & - & y & & & = & 1 \end{array}$$

has infinitely many solutions, because by suming up the first two equations we get the third one. Line p lies in plane ϱ. □

The following exercise is a typical vector spaces intersection exercise. Reader should be able to solve this exercise. We recommend not to continue in reading this book unless it is so.

4.7. Find intersection of 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. Point $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^4$ lies in $Q_1 \cap Q_2$ 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 $t_1, t_2, t_3 \in \mathbb{R}$ and, as well,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix}$$

for some $s_1, s_2 \in \mathbb{R}$. We get an equation

$$t_1 \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \end{pmatrix} + t_2 \begin{pmatrix} 2 \\ 4 \\ 5 \\ 1 \end{pmatrix} + t_3 \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 4-4 \\ 4+5 \\ 4-1 \\ 4+2 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix}.$$

It remains to cope with the consequences of transition from the former coordinate system $(A; \underline{u})$ to the our 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 will be again via system of linear equations, but this time nonhomogeneous in general. □

4.5. 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 in that the origin of the latter one is translated about the vector $(B_0 - A_0)$. Hence we can read off the equations for the corresponding coordinate transformations 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 $y = (y_1, \ldots, y_n)^T$ denotes the column of coordinates of the vector $(A_0 - B_0)$ in the basis $\underline{v}$, and let $M = (a_{ij})$ be the matrix expressing the basis $\underline{u}$ in terms of the basis $\underline{v}$. Then

$$\begin{aligned} x_1' &= y_1 + a_{11}x_1 + \cdots + a_{1n}x_n \\ &\vdots \\ x_n' &= y_n + a_{n1}x_1 + \cdots + a_{nn}x_n \end{aligned}$$

i.e.in matrix notation

$$x' = y + M \cdot x.$$

As an example, we may express the influence of such a change of basis on the coordinates of subsets described by systems of linear equations. Let our 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 will have 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 an one affine frame, then it is so also in the all other affine frames. This finishes fully the proof of the previous proposition.

4.6. 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 an one-dimensional vector space $\mathbb{R}$ (and the supporting set is also $\mathbb{R}$). The affine coordinates are obtained by a 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 exactly 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 then are all points and straight lines in the plane (0-dimensional and 1-dimensional). The lines are prescribed by a

Using matrix notation (variables are t_1, t_2, t_3, s_1, s_2 respectively and we move vectors corresponding to s_1 and s_2 to the left-hand side) we solve by row operations

$$\left(\begin{array}{ccccc|c} 3 & 2 & 0 & 0 & 1 & 0 \\ 5 & 4 & 3 & 6 & 5 & 9 \\ 4 & 5 & 1 & 2 & 3 & 3 \\ 2 & 1 & 2 & 4 & 3 & 6 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 3 & 2 & 0 & 0 & 1 & 0 \\ 0 & 2 & 9 & 18 & 10 & 27 \\ 0 & 7 & 3 & 6 & 5 & 9 \\ 0 & -1 & 6 & 12 & 7 & 18 \end{array}\right) \sim$$

$$\dots \sim \left(\begin{array}{ccccc|c} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right).$$

We can see that $t_1 = t_2 = s_2 = 0$ and for $s_1 = t \in \mathbb{R}$ we have $t_3 = 3 - 2t$. Note that for determination of $Q_1 \cap Q_2$ it is sufficient to know either t_1, t_2, t_3, or s_1, s_2. Let's go back to expression

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix} + s_2 \begin{pmatrix} -1 \\ -5 \\ -3 \\ -3 \end{pmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + t \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix}.$$

Intersection of given subspaces is line ($s = -2t$)

$$[4, 4, 4, 4] + s\,(0, 3, 1, 2)\,, \qquad s \in \mathbb{R}.$$

For checking correctness of our solution we can substitute

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ -5 \\ 1 \\ -2 \end{bmatrix} + t_1 \begin{pmatrix} 3 \\ 5 \\ 4 \\ 2 \end{pmatrix} + t_2 \begin{pmatrix} 2 \\ 4 \\ 5 \\ 1 \end{pmatrix} + t_3 \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix}$$

$$= \begin{bmatrix} 4 \\ -5 \\ 1 \\ -2 \end{bmatrix} + (3 - 2t) \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + t \begin{pmatrix} 0 \\ -6 \\ -2 \\ -4 \end{pmatrix}.$$

□

4.8. Decide whether 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. Arbitrary pair of given points in $\mathbb{R}^3$ defines a vector (see definition of affine space; its coordinates are given respectively by differences of coordinates of two points). The fact that four points lie in the same plane is equivalent to the fact that three vectors given by one chosen point and one of the other three points are linearly dependent. We can choose for example point $[0, 2, 1]$ (regardless of the choice), then we consider vectors $[0, 2, 1] - [-1, 2, 0] = (1, 0, 1)$, $[0, 2, 1] - [-2, 5, 2] = (2, -3, -1)$ and $[0, 2, 1] - [0, 5, 4] = (0, -3, -3)$. We can see that the sum of twice the first vector and the third vector is equal to the second vector and the vectors are linearly dependent (in

choice of a point and one generator of direction, i.e. a vector from the corresponding difference space (so called 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 a choice of an origin and three linearly independent vecors (directions and scales). The proper affine subspaces are then all points, straight lines and planes (0-dimensional, 1-dimensional and 2-dimensional).

(4) The subspace of all solutions of one linear equation $a \cdot x = b$ for an unknown point $[x_1, \dots, x_n] \in \mathcal{A}_n$, known nonzero vector of coefficients $(a_1, \dots, a_n)$ and a scalar $b \in \mathbb{R}$ is an affine subspace of dimension $n-1$ (we also say that the subspace is of codimension 1), i.e. so called *hyperplane in A_n*.

4.7. Affine combinations of points. Let us now introduce an analogue of the linear combination of vectors. Let $A_0, \dots, A_k$ be points in the affine space $\mathcal{A}$. Their affine hull $\langle\{A_0 \dots, A_k\}\rangle$ can be written as

$$\{A_0 + t_1(A_1 - A_0) + \dots + t_k(A_k - A_0);\, t_1, \dots, t_k \in \mathbb{R}\}$$

and in any affine coordinates (i.e. each point A_i is expressed by a column of scalars) we can write the same set as

$$\langle A_0, \dots, A_k\rangle = \{t_0A_0 + t_1A_1 + \dots + t_kA_k;\, t_i \in \mathbb{R}, \sum_{i=0}^{k} t_i = 1\}.$$

Affine combinations of points

In general, by formulae $t_0A_0 + t_1A_1 + \dots + t_kA_k$ with coefficients satisfying $\sum_{i=0}^{k} t_i = 1$ we mean the points $A_0 + \sum_{i=1}^{k} t_i(A_i - A_0)$, and we all them the *affine combinations of points*.

The points $A_0 \dots, A_k$ are in a *general position* if they generate a k-dimensional affine subspace. It is easy to see from our definitions that this happens if and only if for each A_i the vectors arose as differences of this point A_i and all other vectors A_j are linearly independent. We also observe that an assignment of a series of $(\dim \mathcal{A}) + 1$ points in a general position is equivalent to the definition of an affine frame with the origin in the first of them.

4.8. Simplices. For points in an affine space the affine combination is a similar construction as the linear combination for vectors in a vector space. Indeed, the affine subspace generated by points $A_0 \dots, A_k$ is equal to the set of all affine combinations of its generators. We can generalize also the notion "to lie on the line between two points". In the two-dimensional case we can imagine the interior of a triangle. In general we proceed as follows:

k–dimensional simplices

Let $A_0, \dots, A_k$ be $k + 1$ points in a general position in an affine space $\mathcal{A}$. The set $\Delta = \Delta(A_0, \dots, A_k)$ defined as the set of all affine combinations of points A_i with nonnegative coefficients only, i.e.

$$\Delta = \{t_0A_0 + t_1A_1 + \dots + t_kA_k;\, t_i \in [0, 1] \subset \mathbb{R}, \sum_{i=0}^{k} t_i = 1\},$$

is called a k–dimensional *simplex* generated by the points A_i.

other words, rank of matrix constructed by taking these vectors as rows is less than 3; in this case we have matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 2 & -3 & -1 \\ 0 & -3 & -3 \end{pmatrix},$$

which has rank of 2). Hence, given points lie in the same plane. □

4.9. Into how many parts can three planes slice a space ($\mathbb{R}^3$)? Give an example of planes position for every case.

4.10. Decide whether point [2, 1, 0] lies within convex hull of points [0, 2, 1], [1, 0, 1], [3, −2, −1], [−1, 0, 1].

Solution. We form nonhomogeneous linear system, for coefficients t_1, t_2, t_3, t_4, affine combination of given points, which gives the first point (they are determined unambiguously if the points are not coplanar).

$$\begin{pmatrix} 0 & 1 & 3 & -1 \\ 2 & 0 & -2 & 0 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

The last equation says it is an affine combination. Solving it, we obtain $(t_1, t_2, t_3, t_4) = (1, 0, 1/2, -1/2)$, so it is not a convex combination. □

4.11. In $\mathbb{R}^3$ you are given tetrahedron $ABCD$, where $A = [4, 0, 2]$, $B = [-2, -3, 1]$, $C = [1, -1, -3]$, $D = [2, 4, -2]$.

a) Determine its volume.

b) Decide, whether point $X = [0, -3, 0]$ lies inside this tetrahedron.

Solution. a) Volume of a tetrahedron is one sixth of volume of parallelepiped, of which three edges from the point A are $B - A = (-6, -3, -1)$, $C - A = (-3, -1, -5)$ and $D - A = (-2, 4, -4)$ and it is given by absolute value of 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) Given point does not lie inside the tetrahedron. We express X as an affine combination of its vertices (by solving system of four linear equation in four unknowns a, b, c a d given by equality $X = aA + bB + cC + dD$), and we get $X = \frac{1}{4}A + \frac{1}{2}B + \frac{1}{2}C - \frac{1}{4}D$. This means that X does not lie in the tetrahedron, i.e. in convex hull of points A, B, C and D (a, b, c and d would have to be inside inerval $\langle 0, 1 \rangle$). □

The one–dimensional simplex is a *line segment*, the two–dimensional simplex is a *triangle*, the zero–dimensional simplex is a point.

We notice that each k–dimensional simplex has exactly $k + 1$ *faces* which are defined by equations $t_i = 0$, $i = 0, \dots, k$. We see directly from the definition that the faces are also simplices, 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 a general position is equivalent to the parametric description. We work similarly with the parametric description of simplices.

4.9. 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 also the whole line segment $\Delta(A, B)$. We see directly from the definition that each convex set with $k+1$ points in a general position contains also the whole simplex defined by these points (A formal proof is a part of poof of the following proposition).

The 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*

$$\alpha = \{P + t_1 \cdot v_1 + \dots + t_k \cdot v_k;\ t_1, \dots, t_k \in \mathbb{R}, t_k \geq 0\},$$

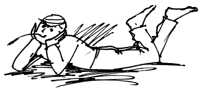
(5) *angles* in two-dimensional subspaces

$$\beta = \{P + t_1 \cdot v_1 + t_2 \cdot v_2;\ t_1 \geq 0, t_2 \geq 0\}.$$

Directly from the definition, it also follows that an intersection of an arbitrary system of convex sets is a convex set. The intersection of all convex sets containing given set M is called the *convex hull* $\mathcal{K}(M)$ of the set M.

Theorem. *The convex hull of any subset $M \subset \mathcal{A}$ is*

$$\mathcal{K}(M) = \{t_1 A_1 + \dots + t_s A_s;\ \sum_{i=1}^{s} t_i = 1,\ t_i \geq 0, A_i \in M\}$$



Proof. Let S denotes the set of all affine combinations on the right-hand side of the equation we want to prove. First we check that S is convex. Therefore, we choose two series of parameters t_i, $i = 1, .., s_1$, t'_j, $j = 1, \dots, s_2$ with the desired properties.

Without loss of generality, we may assume that $s_1 = s_2$ and that the same points from M there appear in both combinations (otherwise we simply add summands with zero coefficients). Let us consider an arbitrary point on the line segment given by vertices defined by the two combinations:

$$\varepsilon(t_1 A_1 + \dots + t_s A_s) + (1 - \varepsilon)(t'_1 A_1 + \dots + t'_s A_s),\ 0 \leq \varepsilon \leq 1.$$

Obviously any point of thise line segment lies in S.

It remains to show that the complex hull of the points $A_1, \dots, A_s$ cannot be smaller than S. The points A_i themselves correspond to the choice of parameters $t_j = 0$ for all $j \neq i$ and $t_i = 1$. Let us assume that the claim holds for all sets with $s - 1$ points at most. It means that the convex hull of the points $A_1, \dots, A_{s-1}$ is (according to the assumption) formed

4.12. Affine transformation of point coordinates

Point X coordinates are expressed as $[2,2,3]$ in affine basis

$\{[1,2,3],(1,1,1),(1,-1,2),(2,1,1)\}$ (in $\mathbb{R}^3$). Determine its coordinates in standard basis, i.e. in basis $\{[0,0,0],(1,0,0),(0,1,0),(0,0,1)\}$.

Solution. Coordinates $[2,2,3]$ in basis $\{[1,2,3],(1,1,1),(1,-1,2),(2,1,1)\}$ give us $[1,2,3]+2\cdot(1,1,1)+2\cdot(1,-1,2)+3\cdot(2,1,1) = [11,5,12]$ coordinates of point X in standard basis. □

4.13. Affine transformation of mapping. Find affine mapping f in coordinate system with basis $\underline{u} = \{(1,1),(-1,1)\}$ and origin $[2,0]$, which is 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 standard basis in $\mathbb{R}^2$.

Solution. Change of basis matrix from basis $\underline{u}$ to the standard basis is

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

We get the transformation matrix in basis $([2,0],\underline{u})$ by first transforming coordinates in basis $([2,0],\underline{u})$ to the standard basis, i.e. to basis $([0,0],(1,0),(0,1))$, then we apply transformation matrix f in the standard basis and in the end we transform back to the coordinates in basis $([2,0],\underline{u})$. Transformation equations for changing coordinates y_1, y_2 in basis $([2,0],\underline{u})$ to coordinates x_1, x_2 in standard basis are

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

And hereby we have

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \end{pmatrix}.\right) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Hence, our sought mapping is

$$\begin{aligned} & f(y_1,y_2) = \\ = & \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}\left[\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}\left(\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix}\right) + \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ = & \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \end{pmatrix} \end{aligned}$$

□

4.14. Let there be a standard coordinate system in $\mathbb{R}^3$ space. Agent K lives at point S with coordinates $[0,1,2]$ and 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 point D with coordinates $[1,1,1]$ and uses coordinate system with basis $\{(0,0,1),(-1,1,2),(1,0,1)\}$. Agent K has set an appointment with agent Bond in the old brickfield which is (according to K's coordinate

exactly by the combinations from the right side of the equation we want to prove, where $t_s = 0$. Now consider a point $A = t_1A_1 + \dots + t_sA_s \in S$, $t_s < 1$, and affine combinations

$$\varepsilon(t_1A_1 + \dots + t_{s-1}A_{s-1}) + (1 - \varepsilon(1-t_s))A_s,\ 0 \le \varepsilon \le \tfrac{1}{1-t_s}.$$

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 $A_1, \dots, A_{s-1}$). The point A is an inner point of this line segment with the parameter $\varepsilon = 1$, and thus A lies in the complex hull of $A_1, \dots, A_s$. □

The convex hulls of finite sets are called *convex polyhedrons*. If and only if the vertices $A_0, \dots, A_k$ defining the convex polyhedron are in a general position, we get a k–dimensional *simplex*. In the case of a simplex, the expression of any of its points as an affine combination of the defining vertices is unique.

A specific example 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

$$\mathcal{P}_k(A; u_1, \dots, u_k) = \{A + c_1u_1 + \dots + c_ku_k;\ 0 \le c_i \le 1\}.$$

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 obvious from the definition that the parallelepipeds are convex. In fact they are the convex hulls of their vertices.

4.10. Examples of standard affine exercises. (1) *To find a parametric description of an implicitly given subspace and vice versa:*

Finding a particular solution of a nonhomogeneous system and a fundamental solution of the homogenized system, we get (in the coordinates in which the equations have been set) exactly the desired parametric description. In the opposite direction, if we write the parametric description in coordinates and then we eliminate the free parameters $t_1, \dots, t_k$, we get exactly 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, e.g. to find a plane in R_3 given by a straight line and a point, by three points etc.), and 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 the parametric form first. Nevertheless, also different methods are advantageous in some concrete situations. Notice that we really need to use one point of each from 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 (and to leave out the linearly dependent). If the system that has arisen is insolvable, then the intersection is empty. In the opposite case, we get an implicit description of the affine subspace which is the intersection we are searching for.

system) at point [1, 1, 0]. Where should Bond come (regarding his coordinate system)?

Solution. Change of basis matrix from agent K's basis to the Bond's one (with the same origins) is

$$T = \begin{pmatrix} -4 & 2 & -1 \\ 1 & 0 & 1 \\ 2 & -1 & 1 \end{pmatrix}$$

Vector $(0, 1, 2)$ thus has coordinates $T \cdot (0, 1, 2)^T = (0, 2, 1)^T$, by transposing origin (we add vector $(-1, 0, 1)$) we get the result $(-1, 2, 2)$. □

4.15. Find a transversal of lines (line passing through both lines)

$$[1, 1, 1] + t(2, 1, 0), \qquad q : [2, 2, 0] + t(1, 1, 1),$$

so that [1, 0, 0] lies on this line.

Solution. We find intersection of sought transversal with line q (denote it by Q). Transversal contains some point lying on p and the point $[1, 0, 0]$, therefore it lies in plane ρ defined by this point and line p, thus in plane

$$[1, 1, 1] + t(2, 1, 0) + s(0, 1, 1).$$

Point Q is then intersection of this plane and line q. We will find it by solving system

$$\begin{aligned} 1 + 2t &= 2 + u \\ 1 + t + s &= 2 + u \\ 1 + s &= u \end{aligned}$$

Left-hand sides of equations represent all three coordinates of an arbitrary point of plane ρ respectively, right-hand sides then represent coordinates of arbitrary point on q (we denoted the free variable as u in order not to be ambiguous). Solving this sytem, we obtain $s = 2$, $t = 2$, $u = 3$ and by inputting $u = 3$ into line q equation we get $Q = [5, 5, 3]$ (we get the same point by inputting $s = 2$, $t = 2$, into parametric equations of ρ). Sought transversal is thereby given by point Q and point $[1, 0, 0]$. Now we easily compute the intersection with p, point $P = [7/3, 5/3, 1]$. □

4.16. Find a common perpendicular of two skew lines

$$\begin{aligned} p : &\quad [3, 0, 3] + (0, 1, 2)t, \quad t \in \mathbb{R} \\ q : &\quad [0, -1, -2] + (1, 2, 3)s \quad s \in \mathbb{R} \end{aligned}$$

Solution. We want to find a transversal perpendicular to both direction vector of line p and direction vector of line q. We can find

If we are given parametric forms, we may also directly search for common points as solutions of the appropriate equations, similarly as we were finding the intersections of vector spaces. In this way, we get directly the parametric description again. If the number of subspaces is greater then two, we must search for the intersection step by step.

If one of the subspaces is defined parametrically and the other implicitly, it suffices to substitute the parametrized coordinates and to solve the resulting system of equations.

(4) *To find a crossbar between 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 a nonempty intersection with both the skew lines. Thus the resulting crossbar is r is an one–dimensional affine subspace. If we are given its one point $A \in r$, then the affine subspace generated by p and A is either a straight line ($A \in p$) or a plane ($A \notin p$). In the first case, we have 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 $r = \langle \{A, B\} \rangle$. The problem has no solution if the intersection is empty. If $q \subset \langle p \cup A \rangle$, we get an infinite number of solutions again, and if the intersection has one element, we get exactly one solution.

If we are given a direction $u \in \mathbb{R}^n$, i.e. the difference space of r, instead of a point, 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 get infinite number of solutions if $q \subset \mathcal{Q}$, otherwise we consider the intersection $\mathcal{Q}$ with q and we finish in the same way as in the previous case.

The solutions of many other practical geometric problems are based mostly on the systematic use of the steps given above.

4.11. Remarks to linear programming. In the beginning of the third chapter in paragraphs 3.4–3.8, we dealt with practical problems which are given by systems of linear inequalities. We easily check that each single inequality

$$a_1x_1 + \cdots + a_nx_n \le b$$

defines a halfspace in the standard affine space $\mathbb{R}^n$ which is bounded by a hyperplane given by the corresponding equation (compare with the definition in paragraph 4.9(4)). Indeed, if we choose the parametric description of the hyperplane

$$\{P + t_1v_1 + \cdots + t_{n-1}v_{n-1}\}$$

with vectors $v_1, \dots, v_{n-1}$ from the difference space, then by completing these vectors by v to a basis of the whole $\mathbb{R}^n$, the value

$$a_1x_1 + \cdots + a_nx_n - b$$

on the linear combination $t_1v_1 + \cdots + t_{n-1}v_{n-1} + t_nv$ must be positive for all vectors with either a positive or a negative t_n.

At the same time we see that the set of all admissible vectors for the problem of the linear programming is always an intersection of a finite number of convex sets and hence the set itself is either convex or empty.

If the intersection is simultaneously nonempty and bounded, then it is obviously a convex polyhedron. As we have justified in 3.4 already, each linear form is either permanently increasing or permanently 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 must have the right direction for example by cross product of those two vectors, and obtain direction $(1, -2, 1)$. Now we form linear equation system which expresses that a vector defined by some two points, one of them lying on p, the other on q, was parallel with direction $(1, -2, 1)$. Symbolically we get system $P - Q = k(1, -2, 1)$, or $\underbrace{[3, 0, 3] + (0, 1, 2)t}_{P} - \underbrace{[0, -1, -2] + (1, 2, 3)s}_{Q} = k(1, -2, 1)$. Treating this equality component-wisely, we get

$$\begin{aligned} 3 - s &= k \\ 1 + t - 2s &= -2k \\ 5 + 2t - 3s &= k \end{aligned}$$

with solution $t = 1$, $s = 2$, $k = 1$. Inputting $t = 1$ into line p parametric equation we get one point of the common perpendicular, point $[3, 1, 5]$, by inputting $s = 2$ into line q equation we then get point $[3, 1, 5]$. The common perpendicular is defined by those two points □

B. Eucledian geometry

4.17. Determine distance of lines in $\mathbb{R}^3$.

$$p : [1, -1, 0] + t(-1, 2, 3), \quad \text{and} \quad q : [2, 5, -1] + t(-1, -2, 1).$$

Solution. The distance is defined as the distance of ortogonal projections of arbitrary points on the respective lines to the ortogonal complement of the vector subspace generated by their directions. We find the ortogonal complement using cross product:

$$\begin{aligned} &\langle(-1, 2, 3), (-1, -2, 1)\rangle^{\perp} = \langle(-1, 2, 3) \times (-1, -2, 1)\rangle \\ &= \langle(8, -2, 4)\rangle = \langle(4, -1, 2)\rangle. \end{aligned}$$

Transversal is for example segment $[1, -1, 0][2, 5, -1]$, so we project vector $[1, -1, 0] - [2, 5, -1] = (-1, -6, 1)$. We obtain distance of lines:

$$\rho(p, q) = \frac{|(-1, -6, 1) \cdot (4, -1, 2)|}{\|(4, -1, 2)\|} = \frac{4}{\sqrt{21}}.$$

□

4.18. Find point A lying on line

$$p : x + 2y + z - 1 = 0, \quad 3x - y + 4z - 29 = 0,$$

which has the same distance from both $B = [3, 11, 4]$ and $C = [-5, -13, -2]$.

Solution. First, we express line p parametrically, we solve system

$$\begin{array}{rcrcrcl} x &+& 2y &+& z &=& 1, \\ 3x &-& y &+& 4z &=& 29. \end{array}$$

the optimal solution in one of the vertices of the corresponding convex polyhedron. The reader should be able to imagine this claim without problems in the case of a two–dimensional or three–dimensional problem. Nevertheless, the straightforward explanation in these small dimensions hold also for all finite–dimensional cases.

So we have given a "geometric proof" of the existence part of the fundamental theorem 3.7. Also we have translated the initial problem into a discrete (i.e. finite) problem of the given cost function in a finite number of points in the space. An example of a practical algorithm for finding and evaluating the corresponding vertices of the convex polyhedron will be given in the chapter about the discrete mathematics yet.

4.12. Affine maps. A map $f : \mathcal{A} \to \mathcal{B}$ between affine spaces is called the *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 a general position.

Then for an arbitrary affine combination of points $t_0 A_0 + \cdots + t_s A_s \in \mathcal{A}$ we get

$$\begin{aligned} f(t_0 A_0 + \cdots + t_s A_s) = & \\ &= f(A_0 + t_1(A_1 - A_0) + \cdots + t_s(A_s - A_0)) \\ &= f(A_0) + t_1 \varphi(A_1 - A_0) + \cdots + t_s \varphi(A_s - A_0) \\ &= t_0 f(A_0) + t_1 f(A_1) + \cdots + t_s f(A_s). \end{aligned}$$

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. Then choosing successively the coefficients $t_0 = 0$ and $t_i = 1$, we define the value of the map φ between difference spaces by the relation $\varphi(A_i - A_0) = f(A_1)$. The previous computation can be read in the opposite direction, and so we can check the correctness and linearity of φ. Indeed, the assumption that the first and the last rows are equal implies that the second and the third rows are equal. So we found out that we really get 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. *The 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. Indeed, the affine combination of $k + 2$ points A_0, 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 $r + s = 1$. Simply we choose a point which is an affine combination of $k + 1$ points only and then we 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.

We rewrite the system as an augmented matrix and perform row operations

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 3 & -1 & 4 & 29 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & -7 & 1 & 26 \end{array}\right) \sim$$

$$\sim \qquad \left(\begin{array}{ccc|c} 1 & 0 & 9/7 & 59/7 \\ 0 & 1 & -1/7 & -26/7 \end{array}\right).$$

We obtain expression

$$p : \left[\frac{59}{7}, -\frac{26}{7}, 0\right] + t\left(-\frac{9}{7}, \frac{1}{7}, 1\right), \quad t \in \mathbb{R}.$$

Introducing substitution $t = 7s + 26$ we get

$$p : [-25, 0, 26] + s\,(-9, 1, 7)\,, \quad s \in \mathbb{R}.$$

We get point A by choosing certain $s \in \mathbb{R}$. On top of that vectors

$$A - B = (-28 - 9s, -11 + s, 22 + 7s)\,,$$
$$A - C = (-20 - 9s, 13 + s, 28 + 7s)$$

should be of the same length, i.e.

$$\sqrt{(-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2}$$
$$= \sqrt{(-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2}\,,$$

or rather

$$(-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2$$
$$= (-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2.$$

should hold. >From the last equation we get $s = -3$. Therefore

$$A = [-25, 0, 26] - 3\,(-9, 1, 7) = [2, -3, 5].$$

□

4.19. Michael is standing in $[2, 1, 2]$ and has a stick of length 4. Can he touch lines p and q with this stick at the same time?

$$p \;:\; [-1, 4, 1] + t(-1, 2, 0),$$
$$q \;:\; [4, 4, -1] + s(1, 2, -4)?$$

(Stick has to pass through $[2, 1, 2]$.)

Solution. We know how to compute transversal of those lines passing through $[2, 1, 2]$. It is segment $[1, 0, 1][3, 2, 3]$, its length is $\sqrt{12}$, which is less than 4. Michael is able to touch the lines. □

4.20. In Euclidian space $\mathbb{R}^4$ determine the distance of point $A = [2, -5, 1, 4]$ and subspace defined by equations

$$U : 4x_1 - 2x_2 - 3x_3 - 2x_4 + 12 = 0, \quad 2x_1 - x_2 - 2x_3 - 2x_4 + 9 = 0.$$

Solution. First, we find a parametric expression of subspace U. For example,

$$B = [0, 3, 0, 3] \in U.$$

4.13. Ratio of colinear points. The affine combinations of pairs of points can be also expressed with the help of so called *ratio of points* on a straight line. If C is given by an affine combination of points A and $B \neq C$ where $C = rA + sB$, 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 the point C as

$$C = A + s(B - A) = B + r(A - B),$$

the ratio λ is the ratio of length of the oriented vectors $C - A$ and $C - B$. In particular, $\lambda = -1$ if and only if C is the center of the line segment between A and B (i.e. $r = s = \frac{1}{2}$ in our affine combination).

Hence our characterization of affine maps in terms of affine combinations has the following intelligible consequence:

Corollary. *Affine maps are exactly those maps which keep the ratios invariant.*

4.14. 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 get the coordinate expression of the affine map $f : \mathcal{A} \to \mathcal{B}$. It follows directly from the definition that it is sufficient to express the image $f(A_0)$ of the origin of coordinate system on $\mathcal{A}$ in the coordinate system on $\mathcal{B}$, i.e. to express the vector $f(A_0) - B_0$ in the basis $\underline{v}$ as a column of coordinates y_0, and 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.

The transformation of affine coordinates corresponds, similarly as in the case of linear maps, 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 can be easily computed by multiplying and adding matrices and vectors. Indeed, changing basis on the domain by a translation w and a matrix M, i.e. the new coordinates may be written in terms of the old ones as

$$x = w + M \cdot x',$$

and changing the basis on the target by a translation z and a matrix N, i.e. the new coordinates may be written in terms of the old ones as

$$y' = z + N \cdot y,$$

for the map given by the translation y_0 and matrix Y in the old bases we directy compute

$$y' = z + N \cdot y = z + N \cdot (y_0 + Y \cdot x)$$
$$= (z + N \cdot y_0 + N \cdot Y \cdot w) + (N \cdot Y \cdot M) \cdot x'.$$

Hence the affine map in the new bases is given by the translation vector $z + N \cdot y_0 + N \cdot Y \cdot w$ a maticí $N \cdot Y \cdot M$.

We know that the distance of A and U equals the length of orthogonal projection of vector $A - B$ to the ortogonal complement of direction of subspace U. However we know the ortogonal complement of 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 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 $a, b \in \mathbb{R}$. Obviously it must hold that $(A-B-P_{A-B}) \perp V$, 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} we obtain

$$((2,-8,1,1) - a(4,-2,-3,-2) - b(2,-1,-2,-2))$$
$$\cdot(4,-2,-3,-2) = 0,$$

$$((2,-8,1,1) - a(4,-2,-3,-2) - b(2,-1,-2,-2))$$
$$\cdot(2,-1,-2,-2)) = 0;$$

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 those dot products, we get system

$$\begin{array}{ccccccc} 19 & - & 33a & - & 20b & = & 0, \\ 8 & - & 20a & - & 13b & = & 0, \end{array}$$

with 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

$$||\,P_{A-B}\,|| = \sqrt{4^2 + (-2)^2 + (-1)^2 + 2^2} = 5.$$

Recall that distance of A and U equals $||\,P_{A-B}\,|| = 5$. □

4.15. Euclidean point spaces. So far we have not needed the notions of distance and length for our geometric considerations. But the length of vectors and the angel between vectors, as we defined them at the end of the third part of the second chapter (see 2.40 and farther), play a significant role in many practical problems. In fact, this additional information refers only to the vectors from the difference space, and so there is not much work to be done:

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 $\|B - A\|$, and will be 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 general *euclidean point space* $\mathcal{E}$ of dimension n we mean an affine space, whose difference space is a real n–dimensional euclidean vector space. The notion of a cartesian coordinate system has the obvious meaning again. Since each choice of such a coordinate system gives an identification of $\mathcal{E}$ with the standard space $\mathcal{E}_n$, we deal, without loss of generality, mainly with the standard euclidean spaces and their subspaces.

From the geometric point of view the simple properties of the scalar product like triangular inequality, Cauchy inequality, Bessel inequality etc., derived in the fourth part of the previous chapter (see 3.25), have very useful consequences:

4.16. Theorem. *For the 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 of the points $A = A_0 + a_1e_1 + \cdots + a_ne_n$, $B = A_0 + b_1e_1 + \cdots + b_ne_n$ 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 minimalizes the distance between A and the points in $\mathcal{Q}$. The distance between A and P is equal to 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 minimalize the distances of points $B \in \mathcal{Q}$ and $A \in \mathcal{R}$. The distance between the points P and Q is equal to the length of the orthogonal projection of the vector $A - B$ into $Z(\mathcal{Q})^\perp$ for arbitrary points $B \in \mathcal{Q}$ and $A \in \mathcal{R}$.

Proof. The first three properties follow directly from the properties of length of vectors in spaces with a scalar product, the fourth one follows directly from the expression of the scalar product in an orthonormal basis.

4.21. In vector space $\mathbb{R}^4$ compute distance v between point $[0, 0, 6, 0]$ and 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 will solve the problem by the least squares method. Let U's generating vectors be columns of matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}$$

and we substitute point $[0, 0, 6, 0]$ by corresponding vector $b = (0, 0, 6, 0)^T$. We will solve $A \cdot x = b$, i.e. linear equation system

$$\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & x_3 & = & 0, \\ & & x_2 & - & x_3 & = & 0, \\ x_1 & + & x_2 & + & 2x_3 & = & 6, \\ x_1 & & & + & 3x_3 & = & 0, \end{array}$$

by least squares method. (Note that the system does not have a solution – the distance would be 0 otherwise.) Let's multiply $A \cdot x = b$ by matrix A^T from the left-hand side. Augmented matrix $A^T \cdot A \cdot x = A^T \cdot b$ then 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 we transform the matrix to the normal form

$$\left(\begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 3 & 6 & 3 & 6 \\ 6 & 3 & 15 & 12 \end{array} \right) \sim \left(\begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 0 & 3 & -3 & 0 \\ 0 & -3 & 3 & 0 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right).$$

We continue with backward elimination

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0 & 3 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right),$$

and see the solution

$$x = (2 - 3t, \; t, \; t)^{\,T}, \quad t \in \mathbb{R}.$$

Note that the existence of infinitely many solutions is caused by third vector generating U, which is redundat because

$$3\,(1, 0, 1, 1) - (2, 1, 1, 0) = (1, -1, 2, 3)\,.$$

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 subspace U, which is the nearest point to $[0, 0, 6, 0]$. The distance is therefore

$$v = ||\,[2, 0, 2, 2] - [0, 0, 6, 0]\,|| = \sqrt{2^2 + 0 + (-4)^2 + 2^2} = 2\sqrt{6}.$$

□

Let us look at 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 $B \in \mathcal{Q}$ since any potential change of the point B would show by adding a vector from $Z(\mathcal{Q})$.

Now let us choose $P = A + (-u_2) = B + u_1 \in \mathcal{Q}$. We get

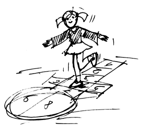
$$\|A - B\|^2 = \|u_1\|^2 + \|u_2\|^2 \geq \|u_2\|^2 = \|A - P\|.$$

From here we see that the minimal possible distance is reached exactly for our point P and its value is $\|u_2\|$ indeed.

We get the general result 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 $u_2 \in (Z(\mathcal{R}) + Z(\mathcal{Q}))^{\perp}$, where the component u_2 does not depend on the choice of the points. Adding suitable vectors from the difference spaces of $\mathcal{R}$ and $\mathcal{Q}$ we obviously obtain points A' and B' whose distance is exactly $\|u_2\|$. □

Now we extend our brief overview of elementary problems in the affine geometry.

4.17. 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 problem is given in the proposition 4.16.

(2) *In $\mathcal{E}_2$ to construct the straight line q through a given point A which form a given angle with a given line p:*

Let us remind that we have worked with angles between vectors in the plane geometry already (see e.g. 2.43). We find a vector $u \in \mathbb{R}^2$ lying in the difference space of the line q, and we 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 two solutions or only one solution.

(3) *To find the perpendicular from a point to a given line:*

The procedure is introduced in the poof of the last but one item of the proposition 4.16.

(4) *In $\mathcal{E}_3$ to determine the distance of two lines p, q:*

We choose arbitrarily one 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 the 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 realizes the minimal possible distance of the given skew lines in terms of the points of intersection. Again, the procedure can be derived from the proof of the proposition 4.16 (the last item). Let η is the subspace generated by a single point $A \in p$ and the sum $Z(p) + (Z(p) + Z(q))^{\perp}$. Provided that the lines p and q are not parallel, it is going to be a plane. Then the intersection $\eta \cap q$ together with the difference space $(Z(p) + Z(q))^{\perp}$ give the parametric expression of the desired axis. If the lines are parallel, then the problem has an infinite number of solutions.

4.18. Angles. Various geometric notions like angles, orientation, volume etc. in the point spaces $\mathcal{E}_n$ are defined in terms of suitable notions from the vector euclidean spaces just as the notion of the distance. Let us remind that we defined the angle between two vectors at the end of the third part of the second chapter, see 2.43.

4.22. Compute volume of parallelepiped in $\mathbb{R}^3$ with base in plane $z = 0$ and with edges given by pairs of vertices $[0,0,0]$, $[-2,3,0]$; $[0,0,0]$, $[4,1,0]$ a $[0,0,0]$, $[5,7,3]$.

Solution. Parallelepiped is given by vectors $(4,1,0)$, $(-2,3,0)$, $(5,7,3)$. We know that its volume is defined as 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 we modified the order of vectors, we would get result ± 42, because determinant gives us *oriented* volume of parallelepiped. Further note that the volume would not change if the third vector was $[a, b, 3]$ for arbitrary $a, b \in \mathbb{R}$. Its surface obviously depends only on ortogonal distance of planes of its upper and lower base and their area

$$\begin{vmatrix} 4 & -2 \\ 1 & 3 \end{vmatrix} = 14.$$

□

4.23. Let points $[0,0,1]$, $[2,1,1]$, $[3,3,1]$, $[1,2,1]$ define a paralleloid. Determine point X lying on line $p : [0,0,1] + (1,1,1)t$ so that parallelepiped defined by given paralleloid and point X has volume of 1.

Solution. We will form a determinant which gives us volume of a parallelepiped with X moving along line p:

$$\begin{vmatrix} t & t & t \\ 2 & 1 & 0 \\ 1 & 2 & 0 \end{vmatrix}.$$

Volume should be 1 which introduces condition $t = 1/3$. □

4.24. 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 angle φ between vectors $F - A$ a $H - A$.

Solution. We have solved this problem using formula for angle between vectors. Let's think about the problem further. Vertices A, F, H are vertices of a triangle with all sides of the same length, it is hence equilateral triangle and therefore $\varphi = \pi/3$. □

4.25. Let S be a midpoint of edge AB of cube $ABCDEFGH$ (with common labelling). Compute cosine of angle between lines ES and BG.

Solution. Dilatation (homotethy) is similar mapping, hence it preserves angles. We can therefore asume that the cube edge has length 1. Further, we can place the point A to the origin of coordinate system and points B and E to points $[1,0,0]$ and $[0,0,1]$ respectively. Other coordinates are then given: $S = [1/2, 0, 0]$, $G = [1,1,1]$, vector

Indeed, from Cauchy inequality follows $0 \le \frac{|u \cdot v|}{\|u\|\|v\|} \le 1$, and so it has 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

$$\cos \varphi(u, v) = \frac{u \cdot v}{\|u\|\|v\|}, \quad 0 \le \varphi(u, v) \le 2\pi.$$

This is completely in accordance with the situation in the two–dimensional euclidean space $\mathbb{R}^2$ and with our philosophy that the notion related to the two vectors is the issue of the plane geometry in fact.

In the euclidean plane, we used also the geometric functions cos and sin which we defined by a pure geometric consideration. We will come back to this in the beginning of the fifth chapter, when we will be able to check precisely the geometric opinion that the function cos is decreasing in the interval $[0, \pi]$. 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), and our defining relation corresponds to the conventions in all dimensions.

In an arbitrary real vector space with a scalar product, it follows directly from definitions that

$$\begin{aligned} \|u - v\|^2 &= \|u\|^2 + \|v\|^2 - 2(u \cdot v) \\ &= \|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \cos \varphi(u, v). \end{aligned}$$

This is evidently the well known *law of cosines* from the plane geometry.

Next, 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 $\|u\|^2$ we get

$$1 = \sum_i (\cos \varphi(u, e_i))^2,$$

which is the law of directional cosines $\varphi(u, e_i)$ of the vector u.

Now we can derive reasonable definitions for angles between general subspaces in an euclidean vector space from the definitions of angles between vectors. Concurrently we must decide how to deal with cases, where the subspaces have a nontrivial intersection. As the angle between two lines, we want to take the smaller one from the two possible angles, in the case of two nonparallel planes in $\mathbb{R}^3$ we do not want to say that the angle is zero since they intersect and have one direction in common:

Angles between subspaces

4.19. Definition. Let us consider finite–dimensional subspaces U_1, U_2 in an euclidean vector space V of an 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 positive and $U_1 \cap U_2 = \{0\}$, then the angle is the minimum of all angles between one–dimensional subspaces

$$\alpha = \min\{\varphi(\langle u \rangle, \langle v \rangle); 0 \ne u \in U_1, 0 \ne v \in U_2\}.$$

$ES = (1/2, 0, -1)$ and $BG = (0, 1, 1)$. Sought cosine of angle φ is then

$$\cos(\varphi) = \left| \frac{(1/2, 0, -1) \cdot (0, 1, 1)}{\|(1/2, 0, -1)\| \, \|(0, 1, 1)\|} \right| = \frac{\sqrt{2}}{\sqrt{5}}$$

. □

4.26. Copmute angle between line p given by implicit equations

$$\begin{array}{rcrcrcl} x & + & 3y & + & z & = & 0, \\ -x & - & y & + & z & = & 0 \end{array}$$

and plane $\varrho : x + y + 2z + 1 = 0$.

Solution. We can see that normal vector of plane ϱ is $(1, 1, 2)$. Now we copy the first equation of line p and then sum both of them, obtaining

$$\begin{array}{rcrcrcl} x & + & 3y & + & z & = & 0, \\ & & 2y & + & 2z & = & 0. \end{array}$$

>From this system we can see that $y = -z$ and $x = 2z$. Vector $(2, -1, 1)$ is therefore direction vector of p; in other words, we have (p is obviously passing through the origin)

$$p : [0, 0, 0] + t\,(2, -1, 1)\,, \quad t \in \mathbb{R}.$$

For angle φ between vectors $(1, 1, 2)$, $(2, -1, 1)$ we have

$$\cos\varphi = \frac{2 - 1 + 2}{\sqrt{6} \cdot \sqrt{6}} = \frac{1}{2}.$$

Hence $\varphi = 60\,^\circ$. However, this is angle between direction vector of p and normal vector ϱ. Sought angle is complement of this angle, so the correct result is $30\,^\circ = 90\,^\circ - 60\,^\circ$. □

4.27. In real plane, find a line which passes through point $[-3, 0]$ and angle between this line and line

$$p : \sqrt{3}x + 3y + 5 = 0$$

is $60\,^\circ$.

Solution. First we have to realize that there will be two such lines. General equation of line in plane has form of

$$ax + by + c = 0, \quad \text{and we may choose parameters so that} \quad a^2 + b^2 = 1.$$

Let's find such numbers $a, b, c \in \mathbb{R}$, so that all the conditions are satisfied. Inputting $x = -3$, $y = 0$ into the equation (line has to go through$[-3, 0]$), we get $c = 3a$. Condition of angle between lines equals $60\,^\circ$ then gives

$$\frac{1}{2} = \cos 60\,^\circ = \frac{\left|\sqrt{3}a + 3b\right|}{\sqrt{12}}, \quad \text{tj.} \quad \sqrt{3} = \left|\sqrt{3}a + 3b\right|.$$

Performing further operations

$$\pm 1 = a + \sqrt{3}b \quad \text{and exponentation} \quad 1 = a^2 + 3b^2 + 2\sqrt{3}ab.$$

We show in a moment that such 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 $U_1 \neq U_1 \cap U_2 \neq U_2$, 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 an euclidean point space $\mathcal{E}_n$ is defined as the angle between their difference spaces $Z(\mathcal{Q}_1)$, $Z(\mathcal{Q}_2)$.

Let us notice that the angle is always well defined, in particular in the last case is

$$(U_1 \cap (U_1 \cap U_2)^\perp) \cap (U_2 \cap (U_1 \cap U_2)^\perp) = \{0\}$$

and so we can indeed determine the angle according to the item (2). Let us also notice that in the case $U_1 \cap U_2 = \{0\}$, the subspaces U_1 and U_2 are perpendicular in terms of our former definitions if and only if the angle between them is $\pi/2$. However, if they have a nontrivial intersection, then they cannot be perpendicular in the former sense.

In order to show the correctness of the definition, it remains to show that the vectors $u \in U_1$, $v \in U_2$ minimalizing the expression for the angle always exist. First a special case:

4.20. Lemma. *Let v be a vector in an euclidean space V and $U \subset V$ an arbitrary subspace. Let us 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. According to the Cauchy inequality, for all vectors $u \in U$ we have

$$\begin{aligned} \frac{|u \cdot v|}{\|u\|\|v\|} &= \frac{|u \cdot (v_1 + v_2)|}{\|u\|\|v\|} = \frac{|u \cdot v_1|}{\|u\|\|v\|} \\ &\leq \frac{\|u\|\|v_1\|}{\|u\|\|v\|} = \frac{\|v_1\|}{\|v\|} = \frac{\|v_1\|^2}{\|v\|\|v_1\|} = \frac{|v_1 \cdot v|}{\|v\|\|v_1\|}. \end{aligned}$$

This implies

$$\cos\varphi(\langle v \rangle, \langle u \rangle) \leq \cos\varphi(\langle v \rangle, \langle v_1 \rangle) = \frac{\|v_1\|}{\|v\|}$$

and thus the vector v_1, which we have found, represents the largest possible value of the cosine of angles between all choices of vectors in U. But since the function cos is decreasing on the interval $[0, \frac{\pi}{2}]$, we get the smallest possible angle in this way, and so the claim is proved. □

4.21. Calculating angles. The procedure in the previous lemma can be understood as follows. We take the orthogonal projection of the one–dimensional subspace generated by v into the subspace U, and we look at the ratio between v and its image. A similar procedure is used in the higher dimension too. However, the problem is to recognize the directions whose projections give the desired (minimal) angle. We can see this in our previous example if we project the bigger space U into one–dimensional $\langle v \rangle$ first, and then orthogonally back to U. We find out that the desired angle corresponds to the direction of

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. □

4.28. Determine general equation of all planes so that angle between every such plane and plane $x+y+z-1=0$ is 60°, and further, they contain line $p:\ [1,0,0]+t\,(1,1,0)$. ○

4.29. Determine angles between 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. Line of intersection between planes has direction vector $(1,-1,1)$, plane ortogonal to this vector has intersection with given planes generated by vectors vektory $(1,0,-1)$ a $(0,1,1)$. Angle between these one-dimensional subspaces is 60°. □

4.30. Cube $ABCDA'B'C'D'$ (in standard notation, i.e. $ABCD$ and $A'B'C'D'$ are faces and AA' is an edge). Compute angle between AB' and AD'.

Solution. Consider cube of side 1 and place it in $\mathbb{R}^3$ in such way that vertex A has coordinates $[0,0,0]$, vertex B coordinates $[1,0,0]$ and vertex C coordinates $[1,1,0]$. Then vertex B' has coordinates $[1,0,1]$ and vertex D' coordinates $[0,1,1]$. We can determine vectors $AB'=B'-A=[1,0,1]-[0,0,0]=(1,0,1)$, $AD'=D'-A=[0,1,1]-[0,0,0]=(0,1,1)$. By definition of angle φ between those vectors

$$\cos(\varphi)=\frac{(1,0,1)\cdot(0,1,1)}{\|\ (1,0,1)\ \|\|\ (0,1,1)\ \|}=\frac{1}{2},$$

hence $\varphi=60°$.

□

For further exercise on angles see .

4.31. Now we will look at usage of Cauchy inequality. Prove that for every $n\in\mathbb{N}$ and for all positive $x_1,x_2,\dots,x_n\in\mathbb{R}$ an inequality

$$n^2\le\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)\cdot(x_1+x_2+\cdots+x_n).$$

holds. For what arguments does equality hold?

the eigenvector of this map, and its eigenvalue is the square of the cosine of the angle.

Hence let us consider two arbitrary subspaces U_1, U_2 in an euclidean vector space V, $U_1\cap U_2=\{0\}$, and let us choose orthonormal bases $\underline{e}$ and $\underline{e}'$ of the whole space V such that $U_1=\langle e_1,\dots,e_k\rangle$, $U_2=\langle e'_1,\dots,e'_l\rangle$.

Let us 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 $(e_1,\dots,e_k)$ and $(e'_1,\dots,e'_l)$, these maps have matrices

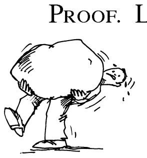
$$A=\begin{pmatrix} e_1\cdot e'_1 & \dots & e_k\cdot e'_1\\ \vdots & & \vdots\\ e_1\cdot e'_l & \dots & e_k\cdot e'_l\end{pmatrix},\ B=\begin{pmatrix} e'_1\cdot e_1 & \dots & e'_l\cdot e_1\\ \vdots & & \vdots\\ e'_1\cdot e_k & \dots & e'_l\cdot e_k\end{pmatrix}.$$

Since we are regarding scalar products on a real vector space, $e_i\cdot e'_j=e'_j\cdot e_i$ holds for all indices i,j, in particular we have $B=A^T$.

The composition of maps $\psi\circ\varphi: U_1\to U_1$ has therefore a symmetric positive semidefinite matrix A^TA, and ψ is an adjoint map to φ. We saw that each such map has only nonnegative real eigenvalues and that it has a diagonal matrix with these eigenvalues on the diagonal in a suitable orthonormal basis, see 3.29 a 3.31.

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^TA. 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. Let us consider all eigenvalues $\lambda_1,\dots,\lambda_k$ (including multiplicities), and let $\underline{u}=(u_1,\dots,u_n)$ be the corresponding orthonormal basis of U_1 made up from the eigenvectors. We may directly 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, i.e. to show that the cosine of the corresponding angle cannot be greater. By the previous lemma, it is sufficient to discuss the angle between u and $\varphi(u)\in U_2$, and we know that $\|u\|=1$. Hence let us choose $v\in U_1$, $v=a_1u_1+\cdots+a_ku_k$, $\sum_{i=1}^k a_i^2=\|v\|^2=1$. Then

$$\begin{aligned}\|\varphi(v)\|^2&=\varphi(v)\cdot\varphi(v)=(\psi\circ\varphi(v))\cdot v\\ &\le\|\psi\circ\varphi(v)\|\|v\|=\|\psi\circ\varphi(v)\|.\end{aligned}$$

Moreover, the previous lemma gives also 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 we have chosen λ_1 to be the largest eigenvalue and the sum of squares of coordinates a_i^2 is fixed to one, we get

$$(\cos\alpha)^2=\|\varphi(v)\|^2\le\|\psi\circ\varphi(v)\|=\sqrt{\blacksquare}\ \sum_{i=1}^{k}(\lambda_i a_i)^2=$$

Solution. It is sufficient to consider Cauchy inequality

$$|\, u \cdot v \,| \le ||\, u \,||\, ||\, v \,||$$

in Euclidean space $\mathbb{R}^n$ for vectors

$$u = \left(\frac{1}{\sqrt{x_1}}, \frac{1}{\sqrt{x_2}}, \dots, \frac{1}{\sqrt{x_n}} \right), \qquad v = \left(\sqrt{x_1}, \sqrt{x_2}, \dots, \sqrt{x_n} \right).$$

We obtain

$$n \le \sqrt{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \cdot \sqrt{x_1 + x_2 + \dots + x_n}. \tag{4.1}$$

We will get wanted inequality by raising (‖4.1‖) to a power. We further know that Cauchy inequality is becoming equality when vector u is a multiple of v, which implies $x_1 = x_2 = \cdots = x_n$. □

4.32. Vectors $\underline{u} = (u_1, u_2, u_3)$ and $\underline{v} = (v_1, v_2, v_3)$ are given. Find third unit vector such that parallelepiped defined by those three vectors had the greatest possible volume.

Solution. Denote sought vector as $\underline{t} = (t_1, t_2, t_3)$. By Proposition ‖**??**‖ is volume of parallelepiped $\mathcal{P}_3(0; \underline{u}, \underline{v}, \underline{t})$ defined as absolute value of determinant

$$\begin{vmatrix} u_1 & v_1 & t_1 \\ u_2 & v_2 & t_2 \\ u_3 & v_3 & t_3 \end{vmatrix} = \begin{vmatrix} t_1 & t_2 & t_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \underline{t} \cdot (\underline{u} \times \underline{v}) \le \|\underline{t}\| \|\underline{u} \times \underline{v}\| = \|\underline{u} \times \underline{v}\|.$$

Sign of inequality follows from Cauchy inequality, moreover we know that 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

$$t = \pm \frac{(\underline{u} \times \underline{v})}{\|(\underline{u} \times \underline{v})\|}.$$

□

4.33. Find foot of line passing through point $[0, 0, 7]$ and perpendicular to plane

$$\rho : \; [0, 5, 3] + (1, 2, 1)t + (-2, 1, 1)s.$$

4.34. In Euclidean space $\mathbb{R}^5$ determine the distance of 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 $t_1, s_1, t_2, s_2 \in \mathbb{R}$, and 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. Case ϱ_1, ϱ_2. We first compute ortogonal complement to sum of vectors defining the planes. We form a matrix where its rows

$$\sqrt{\text{[redacted]}}$$

$$= \quad \lambda_1^2 + \sum_{i=1}^{k} a_i^2(\lambda_i^2 - \lambda_1^2) \le \sqrt{\lambda_1^2}.$$

If $v = u$, we get exactly $\|\varphi(v)\|^2 = \lambda_1^2 \|v\|^2 = \lambda^2$, and thus the angle has the minimal value for this vector. □

4.22. Calculating volume. We met an indication of calculating volumes in the plane geometry already at the end of the fifth part of the first chapter (see 1.34). There we found out that the notion of orientation played the fundamental role. We could imagine the orientation as the decision whether we looked at our plane $\mathbb{R}^2$ from above or from bellow. The difference is in the order of the standard basis vectors e_1 and e_2 on the unite circle. We proceed in the same way in general:

Orientation of a vector space

We say that two bases $\underline{u}$ and $\underline{v}$ of a real vector space V determine the same *orientation* if the transformation matrix between them has a positive determinant. Formally, by the orientation of a vector space V we mean the equivalence class of bases $\underline{u}$ with respect to the equivalence which we defined just now, by the sign of the determinant. Equivalent bases in this sense are called agreeing with the chosen orientation.

It follows directly from the definition that there exist exactly two orientations on every vector space. From each agreeing basis we can obtain a disagreeing one by an arbitrary 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 an euclidean point space whose difference space is oriented. In sequel we consider the standard euclidean space $\mathcal{E}_n$ together with the orientation given by the standard basis of $\mathbb{R}^n$.

Let $u_1, \dots, u_k$ be arbitrary vectors in the difference space $\mathbb{R}^n$, $A \in \mathcal{E}_n$ a point. As an example of a convex set, we defined the parallelepiped $\mathcal{P}_k(A; u_1, \dots, u_k) \subset \mathcal{E}_n$ by

$$\mathcal{P}_k(A; u_1, \dots, u_k) = \{A + c_1 u_1 + \dots + c_k u_k; 0 \le c_i \le 1\}.$$

If the vectors $u_1, \dots, u_k$ are linearly independent, we talk about a k–dimensional parallelepiped $\mathcal{P}_k(A; u_1 \dots, u_k) \subset \mathcal{E}_n$. For given vectors $u_1, \dots, u_k$ we have also the parallelepipeds of the lower dimension

$$\mathcal{P}_1(A; u_1), \dots, \mathcal{P}_k(A; u_1, \dots, u_k)$$

in euclidean subspaces $A + \langle u_1 \rangle, \dots, A + \langle u_1, \dots, u_k \rangle$ at our disposal.

If $u_1, \dots, u_k$ are linearly independent, we define the volume

$$\operatorname{Vol} P_k = 0.$$

Otherwise we think about it as we did in the case of the Gramm–Schmidt orthogonalization

$$\langle u_1, \dots, u_k \rangle = \langle u_1, \dots, u_{k-1} \rangle \oplus \langle u_1, \dots, u_{k-1} \rangle^{\perp} \cap \langle u_1, \dots, u_k \rangle.$$

In this decomposition, u_k is uniquely expressed as

$$u_k = u_k' + e_k$$

where $e_k \perp \langle u_1, \dots, u_{k-1} \rangle$.

are direction vectors of planes. Then we transform this matrix into normal form. We get

$$\begin{pmatrix}1&0&-1&0&0\\0&1&0&0&-1\\1&1&1&0&1\\0&-2&0&0&3\end{pmatrix}\sim\dots\sim\begin{pmatrix}1&0&-1&0&0\\0&1&0&0&-1\\0&0&1&0&1\\0&0&0&0&1\end{pmatrix}.$$

So the ortogonal complement is $\langle(0,0,0,1,0)\rangle$. (It was obvious that vector $(0,0,0,1,0)$ lies within the ortogonal complement. By transforming the matrix into normal form we determined that the ortogonal complement is one-dimensional.) The distance between planes equals the size of perpendicular projection of vector $A_1 - A_2$ into subspace $\langle(0,0,0,1,0)\rangle$ for arbitrary points $A_1 \in \varrho_1$, $A_2 \in \varrho_2$. Choose e.g. $A_1 = [7,2,7,-1,1]$, $A_2 = [2,4,7,-4,2]$. Obviously the ortogonal 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 size of $(0,0,0,3,0)$ gives the sought distance 3.

Case σ_1, σ_2. Sum of directions of σ_1, σ_2 is generated by direction vectors. Denote them by

$$u_1 = (2,1,0,0,1)\,,\quad u_2 = (-2,0,1,1,0)\,,$$
$$v_1 = (2,2,4,0,3)\,,\quad v_2 = (2,0,0,-2,-1)\,.$$

Let's find such points $X_1 \in \sigma_1$, $X_2 \in \sigma_2$, that realize the distance between σ_1 and σ_2. We know that

$$\begin{aligned}X_1 - X_2 = \quad & [0,1,2,0,0] - [3,-1,7,7,3]\\ & +p_1u_1 + q_1u_2 - p_2v_1 - q_2v_2\\ = \quad & (-3,2,-5,-7,-3) + p_1u_1 + q_1u_2 - p_2v_1 - q_2v_2\end{aligned}$$

and it holds that

$$\langle\, X_1 - X_2, u_1\,\rangle = 0, \quad \langle\, X_1 - X_2, u_2\,\rangle = 0,$$
$$\langle\, X_1 - X_2, v_1\,\rangle = 0, \quad \langle\, X_1 - X_2, v_2\,\rangle = 0,$$

tj.

$$\begin{aligned}\langle\,(-3,2,-5,-7,-3), u_1\,\rangle + p_1\,\langle\, u_1, u_1\,\rangle + q_1\,\langle\, u_2, u_1\,\rangle\\ - p_2\,\langle\, v_1, u_1\,\rangle - q_2\,\langle\, v_2, u_1\,\rangle = 0,\end{aligned}$$

$$\begin{aligned}\langle\,(-3,2,-5,-7,-3), u_2\,\rangle + p_1\,\langle\, u_1, u_2\,\rangle + q_1\,\langle\, u_2, u_2\,\rangle\\ - p_2\,\langle\, v_1, u_2\,\rangle - q_2\,\langle\, v_2, u_2\,\rangle = 0,\end{aligned}$$

$$\begin{aligned}\langle\,(-3,2,-5,-7,-3), v_1\,\rangle + p_1\,\langle\, u_1, v_1\,\rangle + q_1\,\langle\, u_2, v_1\,\rangle\\ - p_2\,\langle\, v_1, v_1\,\rangle - q_2\,\langle\, v_2, v_1\,\rangle = 0,\end{aligned}$$

$$\begin{aligned}\langle\,(-3,2,-5,-7,-3), v_2\,\rangle + p_1\,\langle\, u_1, v_2\,\rangle + q_1\,\langle\, u_2, v_2\,\rangle\\ - p_2\,\langle\, v_1, v_2\,\rangle - q_2\,\langle\, v_2, v_2\,\rangle = 0.\end{aligned}$$

By computing those dot products we get linear equation system

We define the absolute value of the volume of a parallelepiped inductively such that we fulfil the idea that it is the product of the volume of the "base" and the "altitude":

$$|\operatorname{Vol}|\mathcal{P}_1(A; u_1) = \|u_1\|$$
$$|\operatorname{Vol}|\mathcal{P}_k(A; u_1, \dots, u_k) = \|e_k\|\,|\operatorname{Vol}|\mathcal{P}_{k-1}(A; u_1, \dots, u_{k-1}).$$

If $u_1, \dots, u_n$ is a basis agreeing with the orientation of V, we define the (oriented) volume of the parallelepiped by

$$\operatorname{Vol}\mathcal{P}_k(A; u_1, \dots, u_n) = |\operatorname{Vol}|\mathcal{P}_k(A; u_1, \dots, u_n),$$

in the case of a nonagreeing basis we set

$$\operatorname{Vol}\mathcal{P}_k(A; u_1, \dots, u_n) = -|\operatorname{Vol}|\mathcal{P}_k(A; u_1, \dots, u_n).$$

The following claim clarifies our former comments that the determinant expresses the volume in a sense. The thing is that the first claim says exactly that we get the volume of the parallelepiped in a k–dimensional space, which is stretched on k vectors, such that we write down their coordinates (in an orthonormal basis) into columns of a matrix and we calculate the determinant.

The formula in the second claim is called *Gramm determinant*. Its advantage is that it is independent on the choice of basis and, therefore, it is better to handle in the case that k is lower then the dimension of the whole space.

Theorem. *Let $\mathcal{Q} \subset \mathcal{E}_n$ be an euclidean subspace, and let $(e_1, \dots, e_k)$ be its orthonormal basis. Then for arbitrary vectors $u_1, \dots, u_k \in Z(\mathcal{Q})$ and $A \in \mathcal{Q}$ the following holds*

(1) $\operatorname{Vol}\mathcal{P}_k(A; u_1, \dots, u_k) = \begin{vmatrix}u_1 \cdot e_1 & \dots & u_k \cdot e_1\\ \vdots & & \vdots\\ u_1 \cdot e_k & \dots & u_k \cdot e_k\end{vmatrix}$

(2) $(\operatorname{Vol}\mathcal{P}_k(A; u_1, \dots, u_k))^2 = \begin{vmatrix}u_1 \cdot u_1 & \dots & u_k \cdot u_1\\ \vdots & & \vdots\\ u_1 \cdot u_k & \dots & u_k \cdot u_k\end{vmatrix}$

Proof. The matrix

$$A = \begin{pmatrix}u_1 \cdot e_1 & \dots & u_k \cdot e_1\\ \vdots & & \vdots\\ u_1 \cdot e_k & \dots & u_k \cdot e_k\end{pmatrix}$$

has the coordinates of vectors $u_1, \dots, u_k$ in the chosen basis in columns, and

$$\begin{aligned}|A|^2 = |A||A| = |A^T||A| = |A^TA|\\ = \begin{vmatrix}u_1 \cdot u_1 & \dots & u_k \cdot u_1\\ \vdots & & \vdots\\ u_1 \cdot u_k & \dots & u_k \cdot u_k\end{vmatrix}.\end{aligned}$$

Hence we see that if (1) holds, then also (2) holds.

The unoriented volume is directly form the definition equal to the product

$$|\operatorname{Vol}|\mathcal{P}_k(A; u_1, \dots, u_k) = \|v_1\|\|v_2\| \dots \|v_k\|,$$

where $v_1 = u_1$, $v_2 = u_2 + a_1^2 v_1, \dots, v_k = u_k + a_1^k v_1 + \dots + a_{k-1}^k v_{k-1}$ is the result of the Gramm-Schmidt orthogonalization.

$$\begin{array}{rcrcrcrcr} 6p_1 & - & 4q_1 & - & 9p_2 & - & 3q_2 & = & 7,\\ -4p_1 & + & 6q_1 & & & + & 6q_2 & = & 6,\\ 9p_1 & & & - & 33p_2 & - & q_2 & = & 31,\\ 3p_1 & - & 6q_1 & - & p_2 & - & 9q_2 & = & -11,\end{array}$$

which we solve by forming matrix and performing elementary row operations.

$$\left(\begin{array}{cccc|c} 6 & -4 & -9 & -3 & 7\\ -4 & 6 & 0 & 6 & 6\\ 9 & 0 & -33 & -1 & 31\\ 3 & -6 & -1 & -9 & -11 \end{array}\right) \sim \dots \sim \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & -1\\ 0 & 0 & 1 & 0 & -1\\ 0 & 0 & 0 & 1 & 2 \end{array}\right).$$

The solutions is $(p_1, q_1, p_2, q_2) = (0, -1, -1, 2)$. We have found

$$X_1 - X_2 = (-3, 2, -5, -7, -3) - u_2 + v_1 - 2v_2 = (-3, 4, -2, -4, 2).$$

The size of vector $(-3, 4, -2, -4, 2)$ and at the same time distance between planes σ_1, σ_2 is hence

$$7 = \sqrt{(-3)^2 + 4^2 + (-2)^2 + (-4)^2 + 2^2}.$$

We determined distance between ϱ_1 and ϱ_2 differently than the distance between σ_1 and σ_2. We could have used both methods in both cases. Let's try the former method for the case of σ_1, σ_2. Let's find ortogonal complement of vector subspace generated by

$$(2, 1, 0, 0, 1)\,, \quad (-2, 0, 1, 1, 0)\,, \quad (2, 2, 4, 0, 3)\,, \quad (2, 0, 0, -2, -1)\,.$$

We get

$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1\\ -2 & 0 & 1 & 1 & 0\\ 2 & 2 & 4 & 0 & 3\\ 2 & 0 & 0 & -2 & -1 \end{pmatrix} \sim \dots \sim \begin{pmatrix} 1 & 0 & 0 & 0 & 3/2\\ 0 & 1 & 0 & 0 & -2\\ 0 & 0 & 1 & 0 & 1\\ 0 & 0 & 0 & 1 & 2 \end{pmatrix},$$

The ortogonal complement is $\langle(-3/2, 2, -1, -2, 1)\rangle$, or rather $\langle(3, -4, 2, 4, -2)\rangle$. Note that distance between σ_1 and σ_2 equals the size of ortogonal projection of vector (difference of arbitrary point in σ_1 and arbitrary point in σ_2)

$$u = (3, -2, 5, 7, 3) = [3, -1, 7, 7, 3] - [0, 1, 2, 0, 0]$$

to this ortogonal complement. Denote the ortogonal 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 it holds

$$\langle\, u - p_u,\; v\,\rangle = 0, \quad \text{tj.} \quad \langle\, u,\; v\,\rangle - a\,\langle\, v,\; v\,\rangle = 0.$$

Computing gives $49 - a \cdot 49 = 0$. Therefore $p_u = 1 \cdot v = v$ and the distance between planes σ_1 and σ_2 is equal

$$||\, p_u\, || = \sqrt{3^2 + (-4)^2 + 2^2 + 4^2 + (-2)^2} = 7.$$

Method of computing distance using ortogonal complement of sum of vector spaces has proven to be „faster way to the solution". With no doubt, it will be the same for planes ϱ_1 a ϱ_2. The second method however reveals points where the distance can be measure (pair

Thus we have

$$(\mathrm{Vol}\,\mathcal{P}_k(A; u_1, \dots, u_k))^2 = \begin{vmatrix} v_1 \cdot v_1 & 0 & \dots & 0\\ \vdots & & \ddots & \\ 0 & 0 & \dots & v_k \cdot v_k \end{vmatrix}$$

$$= \begin{vmatrix} v_1 \cdot v_1 & \dots & v_k \cdot v_1\\ \vdots & & \vdots\\ v_1 \cdot v_k & \dots & v_k \cdot v_k \end{vmatrix}.$$

Let us denote by B the matrix whose columns are formed by the coordinates of vectors $v_1, \dots, v_k$ in the orthonormal basis $\underline{e}$. Since $v_1, \dots, v_k$ have arisen from $u_1, \dots, u_k$ as images under a linear transformation with an upper–triangular matrix C with ones on the diagonal, we have $B = CA$ and $|B| = |C||A| = |A|$. But then $|A|^2 = |B|^2 = |A||A|$, and thus $\mathrm{Vol}\,\mathcal{P}_k(A; u_1, \dots, u_k) = \pm|A|$. The resulting volume is zero if the vectors $u_1, \dots, u_k$ are dependent. Provided that they are independent, the sign of the determinant is positive if and only if the basis $u_1, \dots, u_k$ defines the same orientation as the basis $\underline{e}$. □

We can formulate the following important geometric consequence:

4.23. Corollary. *For each linear map $\varphi : V \to V$ on an euclidean space V,* $\det\varphi$ *is equal to 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 volume equal to* $\det\varphi$*–multiple of of the former volume.*

4.24. Outer product and cross product of vectors. The previous considerations are closely related to so called tensor product of vectors. We will not go farther in this technically more complicated area but we mention at least the case of the outer product $n = \dim V$ of vectors $u_1, \dots, u_n \in V$.

Let $(u_{1j}, \dots, u_{nj})^T$ be coordinate expressions of vectors u_j in a chosen orthonormal basis V, and let M be a matrix with elements (u_{ij}). Then the determinant $|M|$ does not depend on the choice of the basis, and its value is called the *outer product* of the vectors $u_1, \dots, u_n$, and denoted by $[u_1, \dots, u_n]$. Hence the outer product is exactly the oriented product of the corresponding parallelepiped, see 4.22.

Several useful properties of the outer product follow directly from the definition

(1) The map $(u_1, \dots, u_n) \mapsto [u_1, \dots, u_n]$ is antisymmetric n–linear map. It means, it is linear in all arguments, and the interchange of any two arguments causes the change of sign of the result.
(2) The outer product is zero if and only if the vectors $u_1, \dots, u_n$ are linearly dependent.
(3) Vectors $u_1, \dots, u_n$ form a positive basis if and only if their outer product is positive.

In technical applications in the space $\mathbb{R}^3$, we often use a closely related operation, so called cross product, which assigns a vector to any pair of vectors.

Let us consider an arbitrary euclidean vector space V of dimension $n \geq 2$ and vectors $u_1, \dots, u_{n-1} \in V$. If we substitute of points in which the planes are the closest). Let's 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

$$X_1 = [7, 2, 7, -1, 1] + t_1u_1 + s_1u_2,$$
$$X_2 = [2, 4, 7, -4, 2] + t_2v_1 + s_2v_2,$$

so

$$\begin{aligned} X_1 - X_2 = \quad & [7, 2, 7, -1, 1] - [2, 4, 7, -4, 2] \\ & + t_1u_1 + s_1u_2 - t_2v_1 - s_2v_2 \\ = \quad & (5, -2, 0, 3, -1) + t_1u_1 + s_1u_2 - t_2v_1 - s_2v_2. \end{aligned}$$

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 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 only solution $t_1 = -5/2$, $s_1 = 41/2$, $t_2 = 5/2$, $s_2 = -8$. We obtained

$$X_1 = [7, 2, 7, -1, 1] - \frac{5}{2}u_1 + \frac{41}{2}u_2 = \left[\frac{9}{2}, \frac{45}{2}, \frac{19}{2}, -1, -\frac{39}{2}\right],$$
$$X_2 = [2, 4, 7, -4, 2] + \frac{5}{2}v_1 - 8v_2 = \left[\frac{9}{2}, \frac{45}{2}, \frac{19}{2}, -4, -\frac{39}{2}\right].$$

Now we can easily see that the distance between points X_1, X_2 (and, at the same time, distance between planes ϱ_1, ϱ_2) je $||\, X_1 - X_2 \,|| = ||\, (0, 0, 0, 3, 0) \,|| = 3$. □

4.35. Find intersection of plane passing through point $A = [1, 2, 3, 4] \in \mathbb{R}^4$ and ortogonal to plane

$$\varrho : [1, 0, 1, 0] + (1, 2, -1, -2)s + (1, 0, 0, 1)t, \quad s, t \in \mathbb{R}.$$

Solution. First, let's find plane ortogonal to ϱ. Its direction will be ortogonal to direction of ϱ, for vectors (a, b, c, d) within its direction we get linear equation system

$$(a, b, c, d) \cdot (1, 2, -1, -2) = 0 \quad \equiv \quad a + 2b - c - 2d = 0$$
$$(a, b, c, d) \cdot (1, 0, 0, 1) = 0 \quad \equiv \quad a + d = 0.$$

these $n-1$ vectors into the first $n-1$ arguments of the n–linear map defined by the volume determinant as above, then we are given one argument left, i.e. a linear form on V. Since we have the scalar product at disposal, each linear form corresponds to exactly one vector. We call this vector $v \in V$ the *cross product* of vectors $u_1, \dots, u_{n-1}$, i.e. the following holds for each vector $w \in V$

$$\langle v, w\rangle = [u_1, \dots, u_{n-1}, w].$$

We denote the cross product by $v = u_1 \times \dots \times u_{n-1}$.

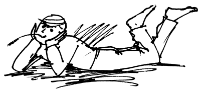
If the coordinates of our vectors in an orthonormal basis are $v = (y_1, \dots, y_n)^T$, $w = (x_1, \dots, x_n)^T$ and $u_j = (u_{1j}, \dots u_{nj})^T$, then our definition can be expressed as

$$y_1x_1 + \dots + y_nx_n = \begin{vmatrix} u_{11} & \dots & u_{1(n-1)} & x_1 \\ \vdots & & \vdots & \vdots \\ u_{n1} & \dots & u_{n(n-1)} & x_n. \end{vmatrix}$$

We see from here that the vector v is given uniquely and its coordinates are calculated by the formal expansion of this determinant along the last column. At the same time, the following properties of the cross product are direct consequences of the definition:

Theorem. *For the cross product $v = u_1 \times \dots \times u_{n-1}$ we have*

(1) $v \in \langle u_1, \dots, u_{n-1}\rangle^{\perp}$
(2) *v is nonzero if and only if the vectors $u_1, \dots, u_{n-1}$ are linearly independent,*
(3) *the length $\|v\|$ of the cross product is equal to the absolute value of the volume of parallelepiped $\mathcal{P}(0; u_1, \dots, u_{n-1})$,*
(4) *$(u_1, \dots, u_{n-1}, v)$ is an agreeing basis of the oriented euclidean space V.*



Proof. The first claim follows directly from the defining formula for v since substituting an arbitrary vector u_j for w we get the scalar product $v \cdot u_j$ on the left and the determinant with two equal columns on the right.

The rank of the matrix with $n-1$ columns u_j is given by the maximal size of a non-zero minor. The minors which define coordinates of the cross product are of degree $n-1$ and thus the claim (2) is proved.

If the vectors $u_1, \dots, u_{n-1}$ are dependent, then also (3) holds. Therefore, let us consider that the vectors are independent, let v be their cross product, and let us choose an orthonormal basis $(e_1, \dots, e_{n-1})$ of the space $\langle u_1, \dots, u_{n-1}\rangle$. It follows from what we have proved that there exists a multiple $(1/\alpha)v$, $0 \neq \alpha \in \mathbb{R}$, such that $(e_1, \dots, e_k, (1/\alpha)v)$ is an orthonormal basis of the whole space V. The coordinates of our vectors in this basis are

$$u_j = (u_{1j}, \dots, u_{(n-1)j}, 0)^T, \quad v = (0, \dots, 0, \alpha)^T.$$

So the outer product $[u_1, \dots, u_{n-1}, v]$ is equal (see the definition of cross product)

$$[u_1, \dots, u_{n-1}, v] = \begin{vmatrix} u_{11} & \dots & u_{1(n-1)} & 0 \\ \vdots & & \vdots & \vdots \\ u_{(n-1)1} & \dots & u_{(n-1)(n-1)} & 0 \\ 0 & \dots & 0 & \alpha \end{vmatrix}$$
$$= \langle v, v\rangle = \alpha^2.$$

Expanding the determinant along the last column we get

$$\alpha^2 = \alpha \operatorname{Vol} \mathcal{P}(0; u_1, \dots, i_{n-1}).$$

Solution is two-dimensional vector space $\langle(0,1,2,0),(-1,0,-3,1)\rangle$. Plane τ ortogonal 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 intersection of planes from both parametric equations. We get 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 only solution (it must be so as matrix columns are linearly independent) $s = -8/19$, $t = 34/19$, $u = -54/19$, $v = -26/19$. Inputting parameter values s and t into parametric form of plane ϱ, we obtain sought intersection $[45/19, -16/19, 11/19, 18/19]$ (needless to say, we get the same solution by inputting the values into τ). □

4.36. Find a line passing through point $[1,2] \in \mathbb{R}^2$ so that angle between this line and line

$$p : [0,1] + t(1,1)$$

is 30°.

Solution. Angle between two lines is angle between their direction vectors. It is sufficient to find direction vector $\underline{v}$ of the line. One way to do so is to rotate direction vector of p by 30°. Rotation matrix for the angle 30° is

$$\begin{pmatrix} \cos 30^\circ & -\sin 30^\circ \\ \sin 30^\circ & \cos 30^\circ \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}.$$

Sought 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

$$[1,2] + \left(\frac{\sqrt{3}}{2} - \frac{1}{2}, \frac{\sqrt{3}}{2} + \frac{1}{2}\right) t.$$

□

4.37. Determine $\cos\alpha$, where α is angle between two adjacent faces of regular octohedron (octohedron has eight equilateral triangles as faces).

Solution. Octohedron is symetric, therefore it does not matter which two faces we choose. Further, without loss of generality, asume octohedron of edge length 1 and place it into standard Cartesian coordinate

Both the remaining two claims from the proposition follow From here. □

4.25. Affine and euclidean properties. Now we can have a think about which properties are related to the affine structure of the space and for which properties we really need the scalar product in the difference space.

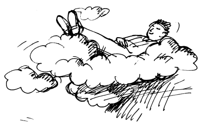
It is obvious that all euclidean transformations, i.e. bijective affine maps between euclidean spaces, which preserve the distance between points preserve also all objects we have studied. I.e. next to the distances they preserve also unoriented angles, unoriented volumes, angle between subspaces etc. If we want them to preserve also oriented angles, cross products, volumes, then we must assume in addition that our transformations preserve the orientation too.

We may formulate our problem also as follows: *Which concepts of euclidean geometry are preserved under affine transformations?*

First let us remind that an affine transformation on a n–dimensional space $\mathcal{A}$ is uniquely defined by mapping $n+1$ points in a general position, i.e. by mapping one n–dimensional simplex. In the plane, it means to choose the image of one (nondegenerate) triangle, which may be an arbitrary (nondegenerate) triangle. The preserved properties will be the properties related to subspaces in particular, i.e. the properties of the type "a line passing through a point" or "a plane contains a line" etc. At the same time, the colinearity of vectors is preserved, and for every two colinear vectors, the ratio of their lengths is preserved (independently on the scalar product defining the length). Similarly, we have already seen that the ratio of volumes of two n–dimensional parallelepipeds is preserved under transformations (since the determinant of the corresponding matrix changes about the same multiple).

These affine properties can be used smartly in the plane to prove geometric claims. 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, and then this property holds for all triangles. Think this argumentation over!

2. Geometry of quadratic forms



After straight lines, the simplest objects in the analytic geometry of plane are so called conic sections. They are given by quadratic equations in cartesian coordinates, and by coefficients we recognize that the conic is a circle, ellipse, parabola or hyperbola, potentially it may be also a pair of lines or a point (the degenerate cases).

We will see that our tools enable us to classify effectively these objects in all finite dimensions and to work with them. It is also obvious that we cannot distinguish a circle from an ellipse in affine geometry, therefore we begin in the euclidean geometry.

4.26. Quadrics in $\mathcal{E}_n$. In analogy with equations of conic sections in plane, we start with objects in euclidean point spaces which are defined in a given orthonormal basis by quadratic equations, we talk about *quadrics*.

system $\mathbb{R}^3$ so that its centroid lies in $[0,0,0]$. Its vertices then are located in points $A = [\frac{\sqrt{2}}{2}, 0, 0]$, $B = [0, \frac{\sqrt{2}}{2}, 0]$, $C = [-\frac{\sqrt{2}}{2}, 0, 0]$, $D = [0, -\frac{\sqrt{2}}{2}, 0]$, $E = [0, 0, -\frac{\sqrt{2}}{2}]$ a $F = [0, 0, \frac{\sqrt{2}}{2}]$.

We will compute angle between faces CDF and BCF. We have to find vectors ortogonal to their intersection and lying within respective faces, which means ortogonal to CF. They are altitudes from D and F to edge CF in triangles CDF and BCF respectively. Altitudes in equilateral triangle are the same segments as medians, so they are SD and SB, where S is midpoint of CF. Because we know coordinates of points C and F, the point S has coordinates $[-\frac{\sqrt{2}}{4}, 0, \frac{\sqrt{2}}{4}]$ and vectors are $SD = (\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})$ a $SB = (\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})$. Together

$$\cos\alpha = \frac{(\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4}) \cdot (\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})}{\|(\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})\| \|(\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})\|} = -\frac{1}{3}.$$

Therefore $\alpha \doteq 132^\circ$. □

4.38. In Euclidean space $\mathbb{R}^5$ determine 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) $U : [0, 1, 1, 0, 0] + t\,(0, 0, 0, 1, -1)\,,\ t \in \mathbb{R},$
$V : [1, 0, 1, 1, 1] + r\,(1, -1, 2, 1, 0) + s\,(0, 1, 3, 2, 0)$
$\quad + p\,(1, 0, 0, 1, 0) + q\,(1, 3, 1, 0, 0)\,,$
$\quad r, s, p, q \in \mathbb{R};$

(e) $U : [0, 2, 5, 0, 0] + t\,(2, 1, 3, 5, 3) + s\,(0, 3, 1, 4, -2)$
$\quad + r\,(1, 2, 4, 0, 3)\,,\ t, s, r \in \mathbb{R},$
$V : [0, 0, 0, 0, 0] + p\,(-1, 1, 1, -5, 0)$
$\quad + q\,(1, 5, 1, 13, -4)\,,\ p, q \in \mathbb{R};$

(f) $U : [1, 1, 1, 1, 1] + t\,(1, 0, 1, 1, 1) + s\,(1, 0, 0, 1, 1)\,,\ t, s \in \mathbb{R},$
$V : [1, 1, 1, 1, 1] + p\,(1, 1, 1, 1, 1) + q\,(1, 1, 0, 1, 1)$
$\quad + r\,(1, 1, 0, 1, 0)\,,$
$p, q, r \in \mathbb{R}.$

Solution. First, recall that angle between affine subspaces is the same as the angle between vector spaces associated to them, and therefore we omit transposition caused by point addition.

Case (a). Since U a V are one-dimensional spaces, angle $\varphi \in [0, \pi/2]$ is given by formula

Let us choose a fixed cartesian coordinate system in $\mathcal{E}_n$ (i.e. a point and an orthonormal basis of the difference space), and let us consider a general quadratic equation for the coordinates $(x_1, \dots, x_n)^T$ of a point $A \in \mathcal{E}_n$

$$\sum_{i,j=1}^{n} a_{ij}x_ix_j + \sum_{i=1}^{n} 2a_ix_i + a = 0, \tag{4.4}$$

where we may assume the symmetry $a_{ij} = a_{ji}$ 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 $a \in \mathbb{R}$. Farther, let us assume that at least one coefficient a_{ij} is nonzero (the equation is linear and describes an euclidean subspace otherwise).

Let us notice that every euclidean (ar affine) coordinate transformation transforms the equation (4.4) into the same form with a quadratic, linear and constant part.

4.27. Quadratic forms. Let us begin our discussion of equation (4.4) with its quadratic part, i.e. bilinear symmetric form $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Similarly, we may 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 $x = x_1e_1 + \dots + x_ne_n$ is given by the equation

$$f(x) = F(x, x) = \sum_{i,j} x_ix_jF(e_i, e_j) = x^T \cdot A \cdot x$$

where $A = (a_{ij})$ is a symmetric matrix with elements $a_{ij} = F(e_i, e_j)$. We call such maps f *quadratic forms*, and the formula 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 we mean the restriction $f(x)$ of a symmetric bilinear form $F(x, y)$ to arguments of the type (x, x). Evidently, we can reconstruct the whole bilinear form F 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

$$f(x) = (S \cdot x')^T \cdot A \cdot (S \cdot x') = (x')^T \cdot (S^T \cdot A \cdot S) \cdot x'.$$

Now let us assume again that our vector space is equipped with a scalar product. Then the previous computation can be formulated as follows. The matrix of 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 (indeed, then we have $S^{-1} = S^T$). We can interpret this result also 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.

$$\cos\varphi = \frac{|\,(1,0,2,-2,1)\cdot(2,0,-2,1,-1)\,|}{||\,(1,0,2,-2,1)\,||\cdot||\,(2,0,-2,1,-1)\,||} = \frac{5}{\sqrt{10}\cdot\sqrt{10}}.$$

Therefore $\cos\varphi = 1/2$ and $\varphi = \pi/3$.

Case (b). We know direction vector $(2, 0, 0, 2, 1)$ of subspace U and normal vector $(1, 1, 1, 0, 1)$ of subspace V. Angle between them $\psi = \pi/3$ can be easily derived from the formula

$$\cos\psi = \frac{(2,0,0,2,1)\cdot(1,1,1,0,1)}{||\,(2,0,0,2,1)\,||\cdot||\,(1,1,1,0,1)\,||} = \frac{3}{3\cdot 2}.$$

Now we have to realise that $\varphi = \pi/2 - \psi = \pi/6$ (because φ is complement to ψ).

Case (c). Hyperplanes U and V are defined by normal vectors $u = (2, -1, 2, 0, 1)$ and $v = (1, 2, 2, 0, 1)$. Obviously angle φ is equal to angle between 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

$$u = (0,0,0,1,-1)\,, \quad v_1 = (1,-1,2,1,0)\,,$$

$$v_2 = (0,1,3,2,0)\,, \quad v_3 = (1,0,0,1,0)\,, \quad v_4 = (1,3,1,0,0)$$

and denote ortogonal projection of u into vector subspace of V (subspace generated by v_1, v_2, v_3, v_4) by p_u. If we knew p_u, from the formula

$$\text{(4.2)} \qquad \cos\varphi = \frac{||\,p_u\,||}{||\,u\,||}$$

would $\varphi \in [0, \pi/2]$. We know that

$$p_u = av_1 + bv_2 + cv_3 + dv_4 \quad \text{for some} a, b, c, d \in \mathbb{R}$$

and that

$$\langle\, p_u - u, v_1 \,\rangle = 0, \quad \langle\, p_u - u, v_2 \,\rangle = 0,$$

$$\langle\, p_u - u, v_3 \,\rangle = 0, \quad \langle\, p_u - u, v_4 \,\rangle = 0.$$

Substituting for p_u we get linear equation system

$$\begin{array}{rcrcrcrcl} 7a & + & 7b & + & 2c & & & = & 1, \\ 7a & + & 14b & + & 2c & + & 6d & = & 2, \\ 2a & + & 2b & + & 2c & + & d & = & 1, \\ & & 6b & + & c & + & 11d & = & 0. \end{array}$$

Solution is $(a, b, c, d) = (-8/19, 7/19, 13/19, -5/19)$, a tak

$$p_u = -\frac{8}{19}v_1 + \frac{7}{19}v_2 + \frac{13}{19}v_3 - \frac{5}{19}v_4 = (0,0,0,1,0)\,,$$

$$\cos\varphi = \frac{||\,(0,0,0,1,0)\,||}{||\,(0,0,0,1,-1)\,||} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}.$$

Hence $\varphi = \pi/4$.

Case (e). Let's determine intersection of vector subspaces associated with given affine subspaces. Vector $(x_1, x_2, x_3, x_4, x_5)$ is in vector subspace of U, if and only if

$$(x_1, x_2, x_3, x_4, x_5) =$$
$$t\,(2,1,3,5,3) + s\,(0,3,1,4,-2) + r\,(1,2,4,0,3)$$

Proof. Indeed, each bilinear form with a fixed second argument becomes a linear form $\alpha_u(\ ) = F(\ , u)$, and in the presence of a scalar product, it must be given by formula $\alpha(u)(v) = v \cdot w$ for a suitable vector w. We set $\varphi(u) = w$. One show directly from the coordinate expression displayed above that φ is a linear map with 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 also a quadratic form by restriction. □

We get immediately the following consequence of this proposition. For each quadratic form f there exists an orthonormal basis of the difference space in which f has a diagonal matrix (and the values on the diagonal are determined uniquely up to their order).

Due to the identification of quadratic forms with linear maps, we can also define correctly the *rank of the quadratic form* as the rank of its matrix in any basis (i.e. the rank is equal to the dimension of the image of the corresponding map φ).

4.28. Classification of quadrics. Let us come back to our equation (4.4). Our results on quadratic forms enable us to rewrite this equation as follows

$$\sum_{i=1}^{n} \lambda_i x_i^2 + \sum_{i=1}^{n} b_i x_i + b = 0.$$

Hence we may assume directly that the quadric is given in this form. In the next step, we do completing the squares for the coordinates x_i with $\lambda_i \neq 0$, which "absorbs" the squares together with the linear terms in the same variable (so called Lagrange algorithm, will be discussed in detail later). So we are left only with linear terms corresponding to variables for which the coefficient at the quadratic term was zero, and we get

$$\sum_{i=1}^{n} \lambda_i (x_i - p_i)^2 + \sum_{j \text{ satisfying } \lambda_j = 0}^{n} b_j x_j + c = 0.$$

This corresponds to a translation of the origin about the vector with coordinates p_i and to such a choice of basis of the difference space that we get the desired diagonal form in the quadratic part. In the identification of quadratic forms with linear maps derived above, it means that φ is diagonal on the orthogonal complement of its kernel. If we are left with some linear terms, we may adjust the orthonormal basis of the difference space for the kernel of φ such that the corresponding linear form is a multiple of the first term of the dual basis. Hence we can already reach the final formula

$$\sum_{i=1}^{k} \lambda_i y_i^2 + b y_{k+1} + c = 0,$$

where k is the rank of matrix of quadratic form f. If $b \neq 0$, we can make the constant c in the equation to be zero by a next change of the origin.

Hence we see that 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.

for some $t, s, r \in \mathbb{R}$, and, at the same time, $(x_1, x_2, x_3, x_4, x_5) \in V$ 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 $p, q \in \mathbb{R}$. Let's find such $t, s, r, p, q \in \mathbb{R}$, so that

$$\begin{aligned} t\,(2, 1, 3, 5, 3) + s\ \ (0, 3, 1, 4, -2) + r\,(1, 2, 4, 0, 3) \\ = p\,(-1, 1, 1, -5, 0) + q\,(1, 5, 1, 13, -4). \end{aligned}$$

It is a homogeneous linear equation system. We will solve it in matrix form (order of variables is t, s, r, p, q)

$$\begin{pmatrix} 2 & 0 & 1 & 1 & -1 \\ 1 & 3 & 2 & -1 & -5 \\ 3 & 1 & 4 & -1 & -1 \\ 5 & 4 & 0 & 5 & -13 \\ 3 & -2 & 3 & 0 & 4 \end{pmatrix} \sim \dots \sim \begin{pmatrix} 1 & 3 & 2 & -1 & -5 \\ 0 & 2 & 1 & -1 & -3 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It has showed that vectors defining V are linear combination of U's vectors. That means V is subset of U, and hence $\varphi = 0$.

Case (f). Again we will find an intersection of U and V. Again we will search for numbers $t, s, p, q, r \in \mathbb{R}$ such that

$$\begin{aligned} &t\,(1, 0, 1, 1, 1) + s\ \ (1, 0, 0, 1, 1) = \\ &p\,(1, 1, 1, 1, 1) + q\,(1, 1, 0, 1, 1) + r\,(1, 1, 0, 1, 0)\,. \end{aligned}$$

The solution is $(t, s, p, q, r) = (-a, a, -a, a, 0), a \in \mathbb{R}$. Intersection $Z(U) \cap Z(V)$ of vector spaces U and V contains exactly vectors

$$\begin{aligned} (0, 0, -a, 0, 0) &= -a\,(1, 0, 1, 1, 1) + a\,(1, 0, 0, 1, 1) \\ &= -a\,(1, 1, 1, 1, 1) + a\,(1, 1, 0, 1, 1) + 0\,(1, 1, 0, 1, 0)\,, \end{aligned}$$

where $a \in \mathbb{R}$, i.e. $Z(U) \cap Z(V)$ is generated by $(0, 0, 1, 0, 0)$ and its ortogonal complement $(Z(U) \cap Z(V))^\perp$ is obviously generated by vectors

$$(1, 0, 0, 0, 0)\,, \quad (0, 1, 0, 0, 0)\,, \quad (0, 0, 0, 1, 0)\,, \quad (0, 0, 0, 0, 1)\,.$$

We get

$$Z(U) \cap Z(V) \neq \{0\}, \quad Z(U) \cap Z(V) \neq Z(U),$$
$$Z(U) \cap Z(V) \neq Z(V).$$

Angle φ is defined as angle between subspaces

$$Z(U) \cap (Z(U) \cap Z(V))^\perp \quad \text{a} \quad Z(V) \cap (Z(U) \cap Z(V))^\perp.$$

It can be further seen that

$$\begin{aligned} Z(U) \cap (Z(U) \cap Z(V))^\perp &= \langle\,(1, 0, 0, 1, 1)\,\rangle\,, \\ Z(V) \cap (Z(U) \cap Z(V))^\perp &= \langle\,(1, 1, 0, 1, 1)\,, (1, 1, 0, 1, 0)\,\rangle\,. \end{aligned}$$

It is enough to express $Z(U)$ as linear combination of vectors

$$(0, 0, 1, 0, 0)\,, \quad (1, 0, 0, 1, 1)$$

and subspace $Z(V)$ by vectors

$$(0, 0, 1, 0, 0)\,, \quad (1, 1, 0, 1, 1)\,, \quad (1, 1, 0, 1, 0)\,.$$

4.29. The case of $\mathcal{E}_2$. As an example of the previous procedure, let us go through the whole discussion in the simplest case of a nontrivial dimension, i.e. dimension two. The original equation has the form

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_1x + a_2y + a = 0.$$

By a suitable choice of a basis of difference space and the subsequent completing the squares we reach the form (we use the same notation x, y for the new coordinates):

$$a_{11}x^2 + a_{22}y^2 + a_1x + a_2y + a = 0$$

where a_i may be nonzero only in the case that a_{ii} is zero. By the last step of the general procedure, i.e. in dimension $n = 2$ only by a choice of a translation, we reach exactly one of the following equations:

$0 = x^2/a^2 + y^2/b^2 + 1$	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

The origin of cartesian coordinates is the *center* of the studied conic, the found orthonormal basis of the difference space gives the direction of *semiaxes*, the final coefficients a, b then give the lengths of semiaxes in nondegenerate directions.

4.30. Affine point of view. In the previous two paragraphs, we have been searching for essential properties and standardized analytical descriptions of objects defined in euclidean spaces by quadratic equations. We wanted to get the simplest equations which may be reached by suitable choice of coordinates. A geometric formulation of our result is that for two different objects – quadrics, given in different cartesian coordinates in general, there exists an *euclidean transformation* on $\mathcal{E}_n$ (i.e. 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, we can obtain directly by our procedure the cartesian coordinates in which our objects are given by the resulting canonical formulas, and hence also the explicit expression of the corresponding coordinate transformation (we know that it is always a composition of a translation, rotation and reflection with respect to a hyperplane).

Of course, we may ask to what extend we can do the same in affine spaces, where we can choose any coordinate system. For example, in the plane it means that we cannot distinguish the circle from the ellipse, but we distinguish from the hyperbola and also between all other types of conics. In particular, all hyperbolas merge into one etc.

We show on quadratic forms the main difference in the procedure, and we postpone the next discussion of this issue to the third part of this chapter.

Let us 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

Since dimension of $Z(U) \cap (Z(U) \cap Z(V))^{\perp}$ is 1, we can use formula ($\|4.2\|$), where $u = (1, 0, 0, 1, 1)$ and p_u is ortogonal projection of u into $Z(V) \cap (Z(U) \cap Z(V))^{\perp}$. Then

$$p_u = a\,(1, 1, 0, 1, 1) + b\,(1, 1, 0, 1, 0)$$

and

$$\langle\, p_u - u, (1, 1, 0, 1, 1)\,\rangle = 0, \quad \langle\, p_u - u, (1, 1, 0, 1, 0)\,\rangle = 0,$$

which leads to linear equation system

$$\begin{array}{rcrcl} 4a & + & 3b & = & 3, \\ 3a & + & 3b & = & 2 \end{array}$$

with only solution $a = 1$, $b = -1/3$. We have computed

$$p_u = \left(\tfrac{2}{3}, \tfrac{2}{3}, 0, \tfrac{2}{3}, 1\right)$$

and from ($\|4.2\|$) it follows that

$$\cos\varphi = \tfrac{\|\,(2/3,2/3,0,2/3,1)\,\|}{\|\,(1,0,0,1,1)\,\|} = \tfrac{\sqrt{7}}{3}, \quad \text{tj.} \quad \varphi \doteq 0,49 \ \ (\approx 28\,°)\,.$$

□

C. Geometry of quadratic forms

4.39. Determine polar basis of form $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = 3x_1^2 + 2x_1x_2 + x_2^2 + 4x_2x_3 + 6x_3^2$.

Solution. Its matrix is

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 1 & 2 \\ 0 & 2 & 6 \end{pmatrix}.$$

According to step (1) of Lagrange algorithm (see Theorem 4.30), we perform following operations

$$\begin{aligned} f(x_1, x_2, x_3) &= \frac{1}{3}(3x_1 + x_2)^2 + \frac{2}{3}x_2^2 + 4x_2x_3 + 6x_3^2 \\ &= \frac{1}{3}y_1^2 + \frac{3}{2}(\frac{2}{3}y_2 + 2y_3)^2 \\ &= \frac{1}{3}z_1^2 + \frac{3}{2}z_2^2 \end{aligned}$$

and we see that the form has rank 2 and matrix changing basis to polar basis $\underline{w}$ is obtained by combination of following transformations:

$$z_3 = y_3 = x_3, \ z_2 = \frac{2}{3}y_2 + 2y_3 = \frac{2}{3}x_2 + 2x_3, \ z_1 = y_1 = 3x_1 + x_2,$$

so the change of basis matrix is

$$T = \begin{pmatrix} 3 & 1 & 0 \\ 0 & \frac{2}{3} & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

We computed polar coordinates, expressed them in standard basis and wrote them as rows of the matrix (see that columns of this matrix are vectors of standard basis in polar basis). Polar basis vector coordinates are the columns of matrix T^{-1}.

V. Then for vector $u = x_1u_1 + \dots + x_nu_n$ we also write the form f as

$$f(x_1, \dots, x_n) = \sum_{ij} a_{ij}x_ix_j,$$

We have already shown in the previous paragraphs with the help of the scalar product that A is diagonal for a suitable choice of basis, i.e. that $F(u_i, u_j) = 0$ for $i \neq j$ holds for a suitable symmetric form F. Each such basis is called the *polar basis* of the quadratic form f. Obviously, we may always choose a scalar product for such purpose. Nevertheless, we are going to prove this statement without use of scalar product, and so we get much simpler algorithm for finding a polar basis among all other basis. At the same time, we get know the relevant information about affine properties of quadratic forms. The following proposition is in literature known as *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.*

Proof. (1) Let A be the matrix of f in basis $\underline{u} = (u_1, \dots, u_n)$ on V, and let us assume $a_{11} \neq 0$. Then we may write

$$\begin{aligned} f(x_1, \dots, x_n) &= a_{11}x_1^2 + 2a_{12}x_1x_2 + \dots + a_{22}x_2^2 + \dots \\ &= a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n)^2 \\ &\quad + \text{terms not containing } x_1. \end{aligned}$$

Hence we transform the coordinates (i.e. we change the basis) such that in new coordinates we have

$$x_1' = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n, \ x_2' = x_2, \dots, x_n' = x_n.$$

It corresponds to the new basis (as an exercise, compute the transformation matrix)

$$v_1 = a_{11}^{-1}u_1, \ v_2 = u_2 - a_{11}^{-1}a_{12}u_1, \dots, v_n = u_n - a_{11}^{-1}a_{1n}u_1$$

and so, as we may expect, in the new basis the corresponding symmetric bilinear form satisfies $g(v_1, v_i) = 0$ for all $i > 0$ (compute!). Thus f has the form $a_{11}^{-1}{x_1'}^2 + h$ in the new coordinates, where h is a quadratic form independent on the variable x_1.

Due to technical reasons, it is mostly better to choose $v_1 = u_1$ in the new basis. Then we have the expression $f = f_1 + h$, where f_1 depend only on x_1', while x_1' does not appear in h, but $g(v_1, v_1) = a_{11}$.

(2) Let us assume that after doing the step (1), we get for h a matrix (of rank less about one) with a nonzero coefficient at ${x_2'}^2$. Then we may repeat exactly the same procedure and we get the expression $f = f_1 + f_2 + h$, where h contains only the variables with index greater than two. We may proceed in this way as long until we get a diagonal form after $n - 1$ steps or in a step, say i-th step, the element a_{ii} is zero.

(3) If the last possibility happens, but in the same time there exists some other element $a_{jj} \neq 0$ with $j > i$, then it suffices to switch the i–th and the j–th vector of the basis and to continue according the the previous procedure.

(4) Let us assume now that we come to the situation $a_{jj} = 0$ for all $j \geq i$. If there is no element $a_{jk} \neq 0$ with $j \geq i, k \geq i$, then we are done since we have got a diagonal matrix. If $a_{jk} \neq 0$, then we use 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 according to (1). □

$$T^{-1} = \begin{pmatrix} \frac{1}{3} & \frac{-1}{2} & 1 \\ 0 & \frac{3}{2} & -3 \\ 0 & 0 & 1 \end{pmatrix},$$

polar basis is therefore $((\frac{1}{3}, 0, 0),\ (-\frac{1}{2}, \frac{3}{2}, 0),\ (1, -3, 1))$. □

4.40. Determine polar basis of form $f : \mathbb{R}^3 \to \mathbb{R}^3$. $f(x_1, x_2, x_3) = 2x_1x_3 + x_2^2$.

Solution. Matrix of the form is

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

We can switch the order of 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. We introduce transformation $z_1 = y_1$, $z_2 = y_2$, $z_3 = y_3 - y_2$. Pak

$$f(x_1, x_2, x_3) = z_1^2 + 2z_2(z_3 + z_2) = z_1^2 + \frac{1}{2}(2z_2 + z_3)^2 - \frac{1}{2}z_3^2.$$

Together we get $z_1 = y_1 = x_2$, $z_2 = y_2 = x_1$, $z_3 = y_3 - y_2 = x_3 - x_1$. Matrix T for change to polar basis is

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad T^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

polar basis is therefore $((0, 1, 0),\ (1, 0, 1)\ (0, 1, 1))$. □

4.41. Find polar basis of quadratic form $f : \mathbb{R}^3 \to \mathbb{R}$, which is in standard basis defined as

$$f(x_1, x_2, x_3) = x_1x_2 + x_1x_3.$$

Solution. By application of Lagrange algorithm we get:

$f(x_1, x_2, x_3) = 2x_1x_2 + x_2x_3$

we perform substitution according to step (4) of the algorithm $y_2 = x_2 - x_1$, $y_1 = x_1$, $y_3 = x_3$

$$= 2x_1(x_1 + y_2) + (x_1 + y_2)x_3 = 2x_1^2 + 2x_1y_2 + x_1x_3 + y_2x_3 =$$

$$= \frac{1}{2}(2x_1 + y_2 + \frac{1}{2}x_3)^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3 =$$

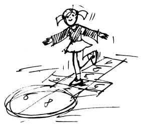
substitution $y_1 = 2x_1 + y_2 + \frac{1}{2}x_3$

$$= \frac{1}{2}y_1^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3 = \frac{1}{2}y_1^2 - 2(\frac{1}{2}y_2 - \frac{1}{2}x_3)^2 + \frac{3}{8}x_3^2 =$$

substitution $y_3 = \frac{1}{2}y_2 - \frac{1}{2}x_3$

$$= \frac{1}{2}y_1^2 - 2y_3^2 + \frac{3}{8}x_3^2.$$

Within coordinates y_1, y_3, x_3 we can see that the quadratic form has a diagonal shape, which means that basis associated with those coordinates is polar basis of the form. If we want to express the basis, we need to get matrix which changes basis from polar to standard. By definition of change of basis matrix, its columns are polar basis vectors.



4.31. Affine classification of quadratic forms. We can improve the Lagrange algorithm for computing polar basis by multiplying the vectors from basis by a scalar such that the coefficients at squares of variables in the corresponding analytic formula for our form will be only 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 our choices in the course of the algorithm. These numbers are called the *signature of a quadratic form*. As before, we get 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* $0 \le p \le r$ *and r independent linear forms* $\varphi_1, \dots, \varphi_r \in V^*$ *such that*

$$f(u) = (\varphi_1(u))^2 + \dots + (\varphi_p(u))^2 - (\varphi_{p+1}(u))^2 - \dots - (\varphi_r(u))^2.$$

Otherwise put, there exists a polar basis, in which f has analytic formula

$$f(x_1, \dots, x_n) = x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_r^2.$$

The number p of positive diagonal coefficients in the matrix of 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 the Lagrange algorithm we obtain $f(x_1, \dots, x_n) = \lambda_1x_1^2 + \dots + \lambda_rx_r^2$, $\lambda_i \neq 0$, in a basis on V. let us assume moreover that exactly the first p coefficients λ_i are positive. Then the transformation $y_1 = \sqrt{\lambda_1}x_1, \dots, y_p = \sqrt{\lambda_p}x_p$, $y_{p+1} = \sqrt{-\lambda_{p+1}}x_{p+1}, \dots, y_r = \sqrt{-\lambda_r}x_r$, $y_{r+1} = x_{r+1}, \dots, y_n = x_n$ yields the desired formula. The forms φ_i are exactly the forms from dual basis in V^* to the polar basis that we obtained. We must prove yet that p does not depend on our procedure. Let us assume that we managed to find a formula for the same form f in the polar bases $\underline{u}$, $\underline{v}$, i.e.

$$f(x_1, \dots, x_n) = x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_r^2$$
$$f(y_1, \dots, y_n) = y_1^2 + \dots + y_q^2 - y_{q+1}^2 - \dots - y_r^2$$

and let us denote the subspace generated by first p vectors of the first basis by $P = \langle u_1, \dots, u_p \rangle$, and similarly $Q = \langle v_{q+1}, \dots, v_n \rangle$. Then for each $u \in P$ we have $f(u) > 0$ while for$v \in Q$ we have $f(v) \le 0$. Hence necessarily $P \cap Q = \{0\}$ holds, and therefore $\dim P + \dim Q \le n$. From here we conclude $p + (n - q) \le n$, i.e. $p \le q$. However, we get also $q \le p$ by the opposite choice of subspaces.

Thus p is independent on the choice of the polar basis. But then for two matrices with the same rank and the same number of positive coefficients in the diagonal form of the corresponding quadratic form, we get the same analytic formulas. □

While we discussed 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 $u \neq 0$,

We get change of basis matrix by either expressing the old variables (x_1, x_2, x_3) by new variables (y_1, y_3, x_3), or equivalently expressing the new ones by the old ones (which is easier), we however need to compute inverse matrix in the latter case.

We have $y_1 = 2x_1 + y_2 + \frac{1}{2}x_3 = 2x_1 + (x_2 - x_1) + \frac{1}{2}x_3$ and $y_3 = \frac{1}{2}y_2 - \frac{1}{2}x_3 = -\frac{1}{2}x_1 + \frac{1}{2}x_3 - \frac{1}{2}x_3$. Matrix changing basis from polar basis to standard basis is

$$T = \begin{pmatrix} 2 & 1 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Inverse matrix is

$$T^{-1} \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{1}{2} \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

One of polar bases of the given quadratic forms is hence for example basis (see the columns of matrix $\{(1/3, 1/3, 0), (-2/3, 4/3, 0), (-1/2, 1/2, 1)\}$. □

4.42. Determine the type of conic section defined by

$$3x_1^2 - 3x_1x_2 + x_2 - 1 = 0.$$

Solution. We complete the squares:

$$\begin{aligned} 3x_1^2 - 3x_1x_2 + x_2 - 1 &= \frac{1}{3}(3x_1 - \frac{3}{2}x_2)^2 - \frac{3}{4}x_2^2 + x_2 - 1 = \\ &= \frac{1}{3}y_1^2 - \frac{4}{3}(\frac{3}{4}x_2 - \frac{1}{2})^2 + \frac{1}{3} - 1 = \\ &= \frac{1}{2}y_1^2 - \frac{4}{3}y_2^2 - \frac{2}{3}. \end{aligned}$$

According to list 4.29, the given conic section is hyperbola. □

4.43. By completing the squares express quadric

$$-x^2 + 3y^2 + z^2 + 6xy - 4z = 0$$

in such way that one can determine its type from it.

Solution. We move all terms containing x to $-x^2$ and complete the square. We get equation

$$-(x - 3y)^2 + 9y^2 + 3y^2 + z^2 - 4z = 0.$$

There are no „unwanted“ terms containing y , so we repeat the procedure for z, which gives us

$$-(x - 3y)^2 + 12y^2 + (z - 2)^2 - 4 = 0.$$

Now we can conclude that there is a transformation of variables that leads to equation (we can divide by 4 first)

$$-\bar{x}^2 + \bar{y}^2 + \bar{z}^2 - 1 = 0.$$

□

(2) *positive semidefinite* if $f(u) \geq 0$ for all vectors $u \in V$,
(3) *negative definite* if $f(u) < 0$ for all vectors $u \neq 0$,
(4) *negative semidefinite* if $f(u) \leq 0$ for all vectors $u \in V$,
(5) *indefinite* if $f(u) > 0$ and $f(v) < 0$ for two vectors $u, v \in V$.

We use the same names also for symmetric matrices corresponding to quadratic forms. By a signature of a symmetric matrix we mean the signature of the corresponding quadratic form.

4.32. 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. We must analyse in detail the form of the transformations used in the Lagrange algorithm for constructing the polar basis. The transformation used in the first step of this algorithm always have an upper triangular matrix T and if we use the technical modification mentioned in the proof of proposition 4.30 moreover, the matrix has one's on the diagonal:

$$T = \begin{pmatrix} 1 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{n2}}{a_{11}} \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \end{pmatrix}.$$

Such matrix of the transformation from basis $\underline{u}$ to basis $\underline{v}$ has several nice properties. In particular, its leading principal submatrices T_k formed by first k rows and columns are the transformation matrices of a subspace $P_k = \langle u_1, \dots, u_k \rangle$ from basis $(u_1, \dots, u_k)$ to basis $(v_1 \dots, v_k)$. The leading principal submatrices A_k of the matrix A of 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 an upper triangular matrix with one's on the diagonal again. Hence we may similarly express A' in terms of A. Thus the determinants of matrices A_k and A'_k are equal by Cauchy formula. So we proved a useful statement:

Let f be a quadratic form on V, $\dim V = n$, *and let $\underline{u}$ be a basis of V such that we never need the items (3) and (4) from the Lagrange algorithm while finding the polar basis. Then as the result we get analytic formula*

$$f(x_1, \dots, x_n) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \dots + \lambda_r x_r^2$$

where r is the rank of form f, $\lambda_1, \dots, \lambda_r \neq 0$ and for leading principal submatrices of the (former) matrix A of quadratic form f we have $|A_k| = \lambda_1\lambda_2 \dots \lambda_k$, $k \leq r$.

In our procedure, each sequential transformation makes zeros under the diagonal in next column. From here it is obvious that if the leading principal minors are nonzero then the next diagonal term in A is nonzero. By this consideration we proved so called *Jacobi theorem*:

Corollary. *Let f be a quadratic form of rank r on a vector space V with matrix A in basis $\underline{u}$. There is no need of other steps in Lagrange algorithm than completing squares if and only if the leading principal submatrices of A satisfy* $|A_1| \neq 0, \dots, |A_r| \neq 0$. *Then*

We can tell the type of the conic section without transforming its equation to the form listed in 4.29. As we know, we can express every conic section as

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0.$$

Determinants $\Delta = \det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{32} & a_{33} \end{vmatrix}$ and $\delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}$ are so called invariants of conic section which means that they are not changed by Euclidian transformation (rotation and translation). Furthermore, different types of conic sections have different signs of those determinants.

- $\Delta \neq 0$ non-degenerate conic sections: ellipse for $\delta > 0$, hyperbola for $\delta < 0$ and parabola for $\delta = 0$ Furthermore, for real ellipse (not imaginary), $(a_{11}+a_{22})\Delta < 0$ must hold.
- $\Delta = 0$ degenerate conic sections, lines

We can easily check that signs (or zero-value) of the determinants are really invariant to coordinate transformation. Denote $X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$ and A is a matrix of quadratic form. Then the corresponding conic section has equation $X^TAX = 0$. We get the standard form by rotation and translation, i.e. by 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 new coordinates $X' = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix}$ holds

$$X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha & c_1 \\ \sin\alpha & \cos\alpha & c_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = MX'. \qquad (4.3)$$

Inputting $X = MX'$ into the conic section equation we get equation in new coordinates

$$\begin{aligned} X^TAX &= 0 \\ (MX')^TA(MX') &= 0 \\ X'^TM^T A\, MX' &= 0. \end{aligned}$$

Denote by A' matrix of the quadratic form in new coordinates. Then $A' = M^TAM$, 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

$$\det A' = \det M^T \det A \det M = \det A = \Delta.$$

there exists a polar basis (which we get by the above algorithm), in which f has analytic formula

$$f(x_1, \dots, x_n) = |A_1|x_1^2 + \frac{|A_2|}{|A_1|}x_2^2 + \dots + \frac{|A_r|}{|A_{r-1}|}x_r^2.$$

Hence if all leading principal minors are positive, then f is positive definite by Jacobi theorem.

On the other hand, let us consider that the form f is positive definite. Then for a suitable regular matrix P we have $A = P^TEP = P^TP$. And so $|A| = |P|^2 > 0$. Let $\underline{u}$ be a chosen basis in which the form f has matrix A. The restrictions of f to subspaces $V_k = \langle u_1, \dots, u_k \rangle$ are positive definite forms f_k again, and the corresponding matrices in bases $u_1, \dots, u_k$ are the leading principal submatrices A_k. Thus $|A_k| > 0$ according to the previous part of the proof.

The claim about negative definite forms follows by observing the fact that A is positive definite if and only if $-A$ is negative definite. □

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 does not have to) intersect. The next reason for finding a more general framework for geometric problems and considerations is to deal only with simple numerical operations like matrix multiplication. Moreover, it is difficult to distinguish very small angles from zero angles, and thus it is preferable to have tools which do not need such distinguishing.

The basic idea of projective geometry is to extend affine spaces by points in infinity such that it allows us an easy work with linear objects like points, lines, planes, projections, etc.

4.33. Projective extension of affine plane. We begin with the simplest interesting case, the geometry in a plane. If we imagine the points in plane $\mathcal{A}_2$ as the plane $z = 1$ in $\mathcal{R}^3$, then each point P in our affine plane is represented by a vector $u = (x, y, 1) \in \mathbb{R}^3$, and so it is represented also by a one–dimensional subspace $\langle u \rangle \subset \mathbb{R}^3$. On the other hand, almost each one–dimensional subspace in $\mathbb{R}^3$ intersects our plane in exactly one point P, and the vectors of such subspace are given by coordinates (x, y, z) uniquely up to a common scalar multiple. Only the subspaces corresponding to vectors $(x, y, 0)$ will not have any intersection with our plane.

Projective plane

Definition. *Projective plane* $\mathcal{P}_2$ is the set of all one–dimensional subspaces in $\mathbb{R}^3$. *Homogeneous coordinates* of point $P = (x : y : z)$ in the projective plane are triples of real numbers given up to a common scalar multiple, while at least one of them must be nonzero. The straight line in projective plane is defined as the set of one–dimensional subspaces (i.e. points in $\mathcal{P}_2$) which generate a two–dimensional subspace (i.e. a plane) in $\mathbb{R}^3$.

Necessarily, also determinant A_{33}, which is algebraic complement of a_{33} is invariant to coordination transformation, because for rotation only $\det A' = \det M^T \det A \det M$ holds. In this case 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.

4.44. Determine type of conic section $2x^2 - 2xy + 3y^2 - x + y - 1 = 0$.

Solution. 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 non-degenerate conic section. Moreover $\delta = 5 > 0$, therefore it is ellipse. Furthermore $(a_{11} + a_{22})\Delta = (2+3) \cdot (-\frac{23}{4}) < 0$, so it is real ellipse. □

4.45. Determine type of conic section $x^2 - 4xy - 5y^2 + 2x + 4y + 3 = 0$.

Solution. 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 hyperbola. □

4.46. Determine equation and type of conic section passing through points

$$[-2, -4], \quad [8, -4], \quad [0, -2], \quad [0, -6], \quad [6, -2].$$

Solution. We will input coordinates of the points into general conic section equation

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_1x + a_2y + a = 0$$

We get linear equation system

$$\begin{array}{rcrcrcrcrcrcl} 4a_{11} & + & 16a_{22} & + & 16a_{12} & - & 2a_1 & - & 4a_2 & + & a & = & 0, \\ 64a_{11} & + & 16a_{22} & - & 64a_{12} & + & 8a_1 & - & 4a_2 & + & a & = & 0, \\ & & 4a_{22} & & & & & - & 2a_2 & + & a & = & 0, \\ & & 36a_{22} & & & & & - & 6a_2 & + & a & = & 0, \\ 36a_{11} & + & 4a_{22} & - & 24a_{12} & + & 6a_1 & - & 2a_2 & + & a & = & 0. \end{array}$$

In matrix form we perform operations

$$\begin{pmatrix} 4 & 16 & 16 & -2 & -4 & 1 \\ 64 & 16 & -64 & 8 & -4 & 1 \\ 0 & 4 & 0 & 0 & -2 & 1 \\ 0 & 36 & 0 & 0 & -6 & 1 \\ 36 & 4 & -24 & 6 & -2 & 1 \end{pmatrix} \sim \cdots$$

In order to have a concrete example, let us look at two parallel lines in affine plane $\mathbb{R}^2$

$$L_1 : y - x - 1 = 0, \quad L_2 : y - x + 1 = 0.$$

If we see the points of lines L_1 and L_2 as finite points in projective space $\mathcal{P}_2$, their homogeneous coordinates $(x : y : z)$ obviously satisfy equations

$$L_1 : y - x - z = 0, \quad L_2 : y - x + z = 0.$$

It is easy to see that the intersection $L_1 \cap L_2$ is the point $(-1 : 1 : 0) \in \mathcal{P}_2$ in this context, i.e. the point of infinity corresponding to the common direction vector of the lines,

4.34. Affine coordinates in projective plane. On the contrary if we begin with the projective plane and if we want to see the affine plane as its "finite" part, then instead of plane $z = 1$ we may take an other plane σ in $\mathbb{R}^3$ which does not pass through origin $0 \in \mathbb{R}^3$. Then the finite points will be those one–dimensional subspaces which have a nonzero intersection with the plane σ.

Let us proceed farther in our example of two parallel lines from the previous paragraph, and let us see what their equations look like in coordinates in affine plane given by $y = 1$. To get them, it suffices to substitute $y = 1$ into the previous equations:

$$L_1' : 1 - x - z = 0, \quad L_2' : 1 - x + z = 0$$

Now the "infinite" points of our former affine plane are given by $z = 0$, and we see that our lines L_1' and L_2' intersects in point $(1, 1, 0)$. This corresponds to the geometric vision that two parallel lines L_1, L_2 in affine plane intersect in infinity, in point $(1 : 1 : 0)$ precisely.

4.35. Projective spaces and transformations. One can generalize in a natural way our procedure from the affine plane to each finite dimension.

Choosing an arbitrary affine hyperplane $\mathcal{A}_n$ in 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 fulfil a hyperplane parallel to $\mathcal{A}_n$, and we call them *infinite points* in the projective extension $\mathcal{P}_n$ of affine plane $\mathcal{A}_n$.

Obviously 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" in infinity and thus the projective line looks like a circle), the projective plane has a projective line of infinite points, the three–dimensional projective space has a projective plane of infinite points etc.

More generally, we define the *projectivization of a vector space*: for an arbitrary vector space V of dimension $n+1$ we define

$$\mathcal{P}(V) = \{P \subset V;\ P \subset V, \dim V = 1\}.$$

Choosing a basis $\underline{u}$ in V we get so called *homogeneous coordinates* on $\mathcal{P}(V)$ such that for a $P \in \mathcal{P}(V)$ we use its arbitrary nonzero vector $u \in V$ and the coordinates of this vector in basis $\underline{u}$. The points of the projective space $\mathcal{P}(V)$ are called *geometric points*, while their generators in V are called *arithmetic representatives*.

In the chosen projective coordinates, we can fix one of them to be one (i.e. we exclude all points of the projective extension which have this coordinate equal to zero), and so we get an embedding

$$\sim \begin{pmatrix} 4 & 16 & 16 & -2 & -4 & 1 \\ 0 & 4 & 0 & 0 & -2 & 1 \\ 0 & 0 & 64 & -8 & 12 & -9 \\ 0 & 0 & 0 & 24 & -36 & 27 \\ 0 & 0 & 0 & 0 & 3 & -2 \end{pmatrix} \sim \cdots$$

$$\sim \begin{pmatrix} 48 & 0 & 0 & 0 & 0 & -1 \\ 0 & 12 & 0 & 0 & 0 & -1 \\ 0 & 0 & 64 & 0 & 0 & 0 \\ 0 & 0 & 0 & 24 & 0 & 3 \\ 0 & 0 & 0 & 0 & 3 & -2 \end{pmatrix}.$$

We can choose value of a. If we choose $a = 48$, we get

$$a_{11} = 1, \quad a_{22} = 4, \quad a_{12} = 0, \quad a_1 = -6, \quad a_2 = 32.$$

Conic section has equation

$$x^2 + 4y^2 - 6x + 32y + 48 = 0.$$

We will complete $x^2 - 6x$, $4y^2 + 32y$ to squares, which gives us

$$(x-3)^2 + 4(y+4)^2 - 25 = 0,$$

or rather

$$\frac{(x-3)^2}{5^2} + \frac{(y+4)^2}{\left(\frac{5}{2}\right)^2} - 1 = 0.$$

We can see it is an ellipse with center in $[3, -4]$. □

4.47. Other characteristics of conic sections. Let's take a further look into some terms related to conic sections. *Axis of conic section* is a line of reflection symmetry for conic section. From canonical form of conic section in polar basis (4.29) it can be derived that an ellipse has two axes ($x = 0$ a $y = 0$), a parabola has one axis ($x = 0$) a hyperbola has two axes ($x = 0$ a $y = 0$). Intersection of axis and conic section itself is called *conic section vertex*. Numbers a, b from canonical form of conic section (which express distance between vertices and origin) are called *semi-axes length*. In the case of ellipse and hyperbola, the axes intersect in the origin. This point is a point of central symmetry for the conic section. This point is called *center of conic section*. Besides vertices and centers there are other interesting points lying on axis of conic section. For ellipse we have *ellipse foci E*, F characterized by property $|EX| + |FX| = 2a$ for arbitrary X lying on ellipse. Following example shows that such points E a F really exist.

4.48. Existence of foci. For ellipse with lengths of semi-axes $a > b$ are points $E = [-e, 0]$ and $F = [e, 0]$, where $e = \sqrt{a^2 - b^2}$ its foci (in polar coordinates).

Solution. Consider points $X = [x, y]$, which satisfy property $|EX| + |FX| = 2a$ and we show that these are exactly ellipse points.

of n–dimensional affine space $\mathcal{A}_n \subset \mathcal{P}(V)$. It is precisely the construction which we used in our example on projective plane.

4.36. Perspective projection. The advantages of projective geometry shows up nicely in the case of perspective projection $\mathbb{R}^3 \to \mathbb{R}^2$. Let us imagine that an observer sitting in the origin observes "one half of the world", i.e. the points $(X, Y, Z) \in \mathbb{R}^3$ with $Z > 0$, and 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:

$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}.$$

It is not only a nonlinear formula but also the accuracy of calculations will be problematic in the case that Z is small.

Extending this transformation to a map $\mathcal{P}_3 \to \mathcal{P}_2$ we get $(X : Y : Z : W) \mapsto (x : y : z) = (-fX : -fY : Z)$, i.e. a map described by simple linear formula

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \\ W \end{pmatrix}$$

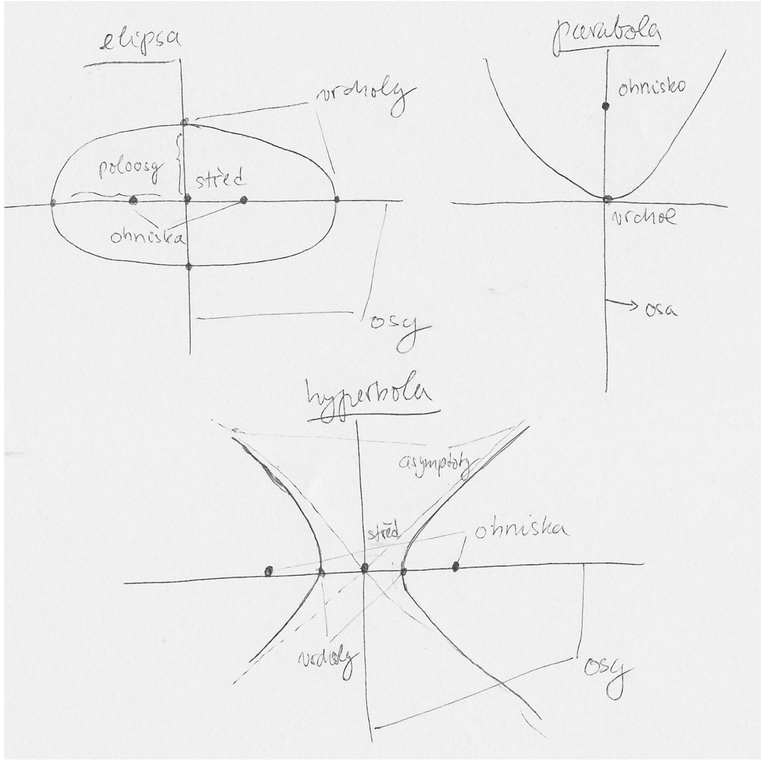
This simple expression defines the perspective projection for finite points in $\mathbb{R}^3 \subset \mathcal{P}_3$ which we substitute as points with $W = 1$. In this way we eliminated problems with points whose image runs to infinity. Indeed, if the Z–coordinate of a real point is close to zero, then the value of the third homogeneous coordinate of the image is close to zero, i.e. it corresponds to a point close to infinity.

4.37. Affine and projective transformations. Obviously, each injective linear map $\varphi : V_1 \to V_2$ between vector spaces maps one–dimensional subspaces to one–dimensional subspaces. Therefore, we get a map on projectivizations $T : \mathcal{P}(V_1) \to \mathcal{P}(V_2)$. Such maps are called *projective maps*, in literature one uses also the notion *collineation* 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 domain and image it is given by multiplication by a matrix. More generally if our auxiliary linear map is not injective, then we define the projective map only outside of its kernel, i.e. on points whose homogeneous coordinates do not map to zero.

Since injective maps $V \to V$ of a vector space to itself are invertible, all projective maps of projective space $\mathcal{P}_n$ to itself are invertible too. 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 they differ by a constant multiple.

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 want to switch to affine coordinates of finite points, i.e we fix the first coordinate to be one, the first element in the first row must be also equal to one. Hence the matrices of collineations



Coordinate-wise, this equation looks like

$$\sqrt{(x+e)^2+y^2}+\sqrt{(x-e)^2+y^2}=2a$$

By raising to a power and performing some operations we get

$$(a^2-e^2)x^2+a^2y^2=a^2(a^2-e^2).$$

Substituting $e^2=a^2-b^2$ and dividing by a^2b^2 we get

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1.$$

□

Remark. Number e from the previous example is called *eccentricity* of an ellipse. Similarly, *hyperbola foci* are points E, F, which satisfy $||EX|-|FX||=2a$ for arbitrary X on hyperbola. You can check that there are two points satisfying this condition $[-e,0]$ and $[e,0]$ (in polar basis), where $e=\sqrt{a^2+b^2}$. *Parabola focus* is a point F of coordinates $F=[0,\frac{p}{2}]$ and it is characterized by a fact that distance between this point and arbitrary X on parabola is equal to the distance between X and line $y=-\frac{p}{2}$.

4.49. Find foci of ellipse $x^2+2y^2=2$.

Solution. We can see from the equation that semi-axes lengths are $a=1$ and $b=\frac{1}{\sqrt{2}}$. We easily compute (see ||4.48||): $e=\sqrt{a^2-b^2}=1$, foci coordinates then are $[-1,0]$ a $[1,0]$. □

preserving finite points of our 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 $b=(b_1,\dots,b_n)^T\in\mathbb{R}^n$ and $A=(a_{ij})$ is an invertible matrix of dimension n. The action of such matrix on vector $(1,x_1,\dots,x_n)$ is exactly a general affine transformation, where b is the translation and A its linear part. Thus the affine maps are exactly those collineations which preserve the hyperplane of infinity points.

4.38. 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 it corresponds to a suitable choice of an image of a suitable arithmetic basis of the vector space V.

But it does not hold in general that the image of an arithmetic basis of V determines the collineation uniquely. We show the core of the problem on a simple example of affine plane. If we choose four points A, B, C, D in the plane such that each three of them are in a general position (i.e. no three of them lie on a line), then we may choose their images in the collineation as follows:

Let us choose arbitrarily their four images A', B', C', D' with the same property, and let us choose their homogeneous coordinates u, v, w, z, u', v', w', z' v $\mathbb{R}^3$. Obviously the vectors z and z' can be written as linear combinations

$$z=c_1u+c_2v+c_3w,\quad z'=c_1'u'+c_2'v'+c_3'w',$$

where all six coefficients must be nonzero, otherwise there exist three of our points which are not in general position.

Now we choose new arithmetic representatives $\tilde{u}=c_1u$, $\tilde{v}=c_2v$ and $\tilde{w}=c_3w$ of points A, B and C respectively, and similarly $\tilde{u}'=c_1u'$, $\tilde{v}'=c_2v'$ and $\tilde{w}'=c_3w'$ for points A', B' a C'. This choice defines an unique linear map φ which maps successively

$$\varphi(\tilde{u})=\tilde{u}',\quad \varphi(v')=\tilde{v}',\quad \varphi(\tilde{w})=\tilde{w}'.$$

But at the same time we have

$$\varphi(z)=\varphi(\tilde{u}+\tilde{v}+\tilde{w})=\tilde{u}'+\tilde{v}'+\tilde{w}'=z'$$

and so indeed the constructed collineation maps the points such as we have chosen in advance. The linear map φ is given uniquely by our construction, thus the collineation is given uniquely by our choice.

Our argumentation 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 on regular collineations of a projective line, these are defined by pairwise different images of three pairwise different points.

The procedure which we used obviously 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.*

4.50. Prove that product of distances between ellipse foci and its arbitrary tangent is constant and tell the value of this constant.

Solution. Consider polar basis. Ellipse matrix has diagonal shape diag$(\frac{1}{a^2}, \frac{1}{b^2}, -1)$ and tangent equation in X=$[x_0, y_0]$ is $\frac{x_0}{a^2}x + \frac{y_0}{b^2}y = 1$. Distance between E,F= $[\mp e, 0]$ and this line is equal

$$\frac{1 \pm e\frac{x_0}{a^2}}{\sqrt{\frac{x_0^2}{a^4} + \frac{y_0^2}{b^4}}}$$

and its product

$$\frac{1 - e^2\frac{x_0^2}{a^4}}{\frac{x_0^2}{a^4} + \frac{y_0^2}{b^4}}$$

If we substitute $e^2 = a^2 - b^2$ and $\frac{y_0^2}{b^2} = 1 - \frac{x_0^2}{a^2}$ (point X is lying on ellipse), we will find out that the previous term is equal to b^2. □

4.51. What are the lengths of semi-axes, when the sum of their lengths equals distance between foci equals 1.

Solution. We solve system

$$\begin{array}{rcl} a+b & = & 1 \\ 2e = \ 2\sqrt{a^2-b^2} & = & 1 \end{array}$$

and find solution $a = \frac{5}{8}$, $b = \frac{3}{8}$. □

4.52. For what slopes k are lines passing through $[-4, 2]$ secant and tangent lines of ellipse defined by

$$\frac{x^2}{9} + \frac{y^2}{4} = 1$$

Solution. Direction vector of the line is $(1, k)$ and its parametric equations then are $x = -4+t$, $y = 2+kt$. Intersection with ellipse satisfies

$$\frac{(-4+t)^2}{9} + \frac{(2+kt)^2}{4} = 1$$

This quadratic equation has discriminant equal to

$$D = -\frac{k}{9}(7k+16)$$

Which means that for $k \in (-\frac{16}{7}, 0)$ there are two solutions (line is secant) and for $k = -\frac{16}{7}$ and $k = 0$ only one solution (line is tangent). □

4.53. Find line tangent to ellipse $3x^2 + 7y^2 = 30$, so that its distance from the center of the ellipse is 3.

Solution. Center of ellipse is in the coordinate system origin and for distance d between line $ax+by+c = 0$ and origin we have $d = \frac{|c|}{\sqrt{a^2+b^2}}$. Tangent then satisfies $a^2 + b^2 = \frac{c^2}{9}$. Equation of tangent passing

Proof. The proof is exactly the same as in dimension two. We recommend to write it in a detail as an exercise. □

4.39. Cross-ratio. Let us remind 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 $\lambda = (C; A, B) = -\frac{s}{r}$. But it is obvious that for example for central projection the ratios are not preserved. Moreover, even the relative position of points on a line does not need to be preserved. On contrary we know from above that we may determine uniquely a projective transformation by choosing arbitrarily images of three pairwise different points on a projective line. But one can show relatively easily that the ratio of such ratios for two distinct points C is preserved:

Let us consider four distinct points A,B, C, D in projective space with arithmetic coordinates x, y, w, z respectively which lie on a projective line. Since these four vectors lie in the subspace generated by $\langle x, y\rangle$, we may write w and z as linear combinations

$$w = t_1x + s_1y, \quad z = t_2x + s_2y$$

and we define so called *cross-ratio of four points* (A, B, C, D) as

$$\rho = \frac{s_1}{t_1}\frac{t_2}{s_2}.$$

The definition is correct since although the vectors x and y are determined up to a scalar multiple, these multiples cancel out in our definition.

Similarly, it is obvious from our definition that each projective transformation preserves cross-ratios. Indeed, if it is given in our arithmetic coordinates by a matrix A, we get images $A \cdot w = t_1A \cdot x + t_2A \cdot y$ and similarly for Az, and therefore the four images will have the same cross-ratio.

Let us discuss the characterization of projective transformations yet. It holds again that they are exactly those maps which preserve cross-ratios. But this is not very practical characterization since it contains implicitly also the claim that these maps must map projective lines to projective lines.

But one can prove much stronger statement which says that 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 $\mathbb{R}^n$. And thus these transformations evidently preserve also cross-ratios.

4.40. Duality. The projective hyperplanes in n–dimensional projective space $\mathcal{P}(V)$ are defined as the projectivizations of n–dimensional vector subspaces i vector space V. Hence in homogeneous coordinates they are defined as kernels of linear forms $\alpha \in V^*$ which are determined again 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 in the same time 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 space the *dual projective space* and we talk about a duality between points and hyperplanes.

through point $[x_T, y_T]$ is $3xx_T + 7yy_T - 30 = 0$. For coordinates of touch point we have

$$\begin{aligned}(3x_T)^2 + (7y_T)^2 &= 100\\ 3x_T^2 + 7y_T^2 &= 30\end{aligned}$$

Its solution is $x_T = \pm\sqrt{\frac{55}{6}}$, $y_T = \pm\sqrt{\frac{5}{14}}$. Considering symmetry of ellipse, we get four solutions $\pm 3\sqrt{\frac{55}{6}}x \pm 7\sqrt{\frac{5}{14}}y - 30 = 0$. □

4.54. Hyperbola $x^2 - y^2 = 2$ is given. Find equation of hyperbola having the same foci and passing through point $[-2, 3]$.

Solution. Eccentricity of given hyperbola is $e = \sqrt{2+2} = 2$. Equation of wanted hyperbola will be $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ and its eccentricity will satisfy $e^2 = a^2 + b^2 = 4$. Condition of point $[-2, 3]$ lying on hyperbola gives $\frac{4}{a^2} - \frac{9}{b^2} = 1$. Respective solutions are $a^2 = 1$, $b^2 = 3$. Sought hyperbola is $x^2 - \frac{y^2}{3} = 1$. □

4.55. Determine equations of hyperbola $4x^2 - 9y^2 = 1$ tangents, which are perpendicular to line $x - 2y + 7 = 0$.

Solution. All lines perpendicular to given line have equation $2x + y + c = 0$ for some c. So the line should have exactly one intersection with given hyperbola, so equation $4x^2 - 9(-2x - c)^2 = 1$ should have one solution. That happens for $D = (36c)^2 - 4.32.(9c^2 + 1) = 0$. Hence $c = \pm\frac{2\sqrt{2}}{3}$. □

4.56. Projective approach to conic section. The term of projective space gives us ability to approach the conic section from new perspective (compare with 4.42). We can understand conic section in $\mathcal{E}_2$ defined by 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 set of points in projective plane $\mathcal{P}^2$ with homogenous coordinates $(x : y : z)$, which are zero points of 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 column vector in coordinates (x, y, z) and matrix A is symmetric matrix (a_{ij}). By Theorem 4.31 there exists a basis, in which this quadratic form has one of the following equation

$$f(x, y, z) = x^2 + y^2 + z^2, \quad f(x, y, z) = x^2 + y^2 - z^2.$$

In the former case there is only solution of $f(x, y, z) = 0$ and therefore the original form does not represent any real conic section. The second quadratic form represents a cone in $\mathbb{R}^3$. We get the corresponding conic section by moving back to inhomogeneous coordinates. That means intersecting the cone with plane which had equation $z = 1$ in the original basis. We immediately get the conic section classification

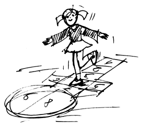
On forms, the linear map defining a given collineation acts by the multiplication of row vectors from right by the same matrix

$$\alpha = (\alpha_0, \dots, \alpha_n) \mapsto \alpha \cdot A,$$

i.e. the matrix of dual maps is A^T. But the dual map maps forms in the opposite direction, from the "target space" to the "initial one", and therefore we need the inverse map to collineation f in order to study the effect of regular collineations on points and their dual hyperplanes. The inverse is given by matrix A^{-1}. Hence the matrix for the action of corresponding collineation on forms is $(A^T)^{-1}$. Since the inverse matrix is equal to the algebraically adjoint matrix A^*_{alg}, up to the multiplication by the inverse of determinant, see equation (2.2) on page 91, we can work directly with the projective transformation of space $\mathcal{P}(V*)$ given by matrix $(A^*_{\text{alg}})^T$ (or without transposing if we multiply row vectors from right).

It is easy to see from definitions that the projective point X belongs to hyperplane α if their arithmetic coordinates satisfy $\alpha \cdot x = 0$. Obviously, it still holds after acting with an arbitrary collineation since

$$(\alpha \cdot A^{-1}) \cdot (A \cdot x) = \alpha \cdot x = 0.$$



4.41. Fixed points, centers and axes. Let us 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 collineation f we mean a point A which is mapped on itself, i.e. $f(A) = A$, by the *fixed hyperplane* of collineation f we mean a hyperplane α which is mapped on itself, i.e. $f(\alpha) \subset \alpha$.

Hence we see directly from the definition that the arithmetic representatives of fixed points are exactly eigenvectors of matrix A.

In the geometry of plane, we met 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 $S \in \mathbb{R}^3$.

Let us note that there appeared also 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 the translation the infinite points behave similarly.

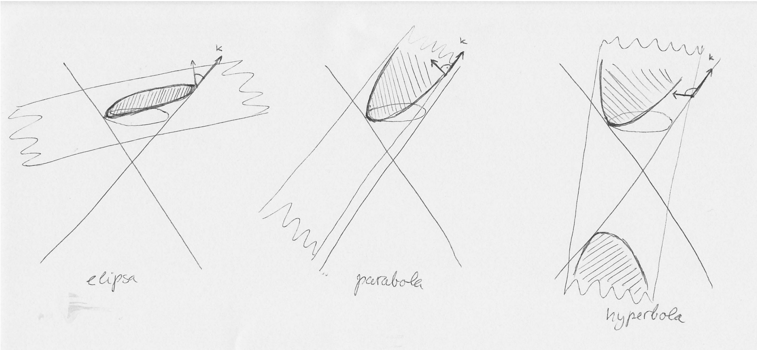
Now we discuss this phenomenon in an arbitrary dimension. First we define a classical notion related to the incidence of points and hyperplanes.

A *bunch of hyperplanes passing through point* $A \in \mathcal{P}(V)$ is the set of all hyperplanes which contain point A. It is obvious from the definition that for each point A the corresponding bunch 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 $f : \mathcal{P}(V) \to \mathcal{P}(V)$ we call a point $S \in \mathcal{P}(V)$ the *center of collineation* f if all hyperplanes in the bunch 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 directly from the definition that the axis of a collineation is the center of the dual collineation, while the bunch 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 (eigenvectors are column respectively row vectors corresponding



from 4.29., which corresponds to intersecting cone in $\mathbb{R}^3$ with different planes. Non-degenerate sections are depicted. Degenerate sections are those which passes the vertex of the cone.

We define following useful terms for 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 $p^T A q = 0$ holds.

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 conic section does not have maximal rank and therefore does define degenerate conic section. Non-degenerate conic sections do not contain singular points.

We call the set of all points X= $\langle x\rangle$ polar conjugated with $P = \langle p\rangle$ *polar* of point P with respect to conic section f. It is therefore set of point for which $F(p, x) = p^T Ax = 0$. Because polar is given by linear combination of coordinates, it is always (in non-singular case) a line. The following part explains geometric interpretation of polar.

4.57. Polar characterization. Consider non-degenerate conic section f. Polar of point $P \in f$ with respect to f is tangent to f with the touch point P. Polar of point $P \notin f$ is line defined by touch points of tangents to f passing through P.

Solution. We will first consider P$\in$ f and show by contradiction that polar has exactly one common point with the conic section (the touch point). Suppose that polar of P, defined by $F(p, x) = 0$, intersects f in Q= $\langle q\rangle$ $\neq$P. Then obviously $F(p, q) = 0$ and $f(q) = F(q, q) = 0$. For arbitrary point $X = \langle x\rangle$ lying on P and Q we then have $x = \alpha p + \beta q$ for some $\alpha, \beta \in \mathbb{R}$. Because of bilinearity and symmetry of F we get

$$f(x) = F(x, x) = \alpha^2 F(p, p) + 2\alpha\beta F(p, q) + \beta^2 F(q, q) = 0$$

to the same eigenvalues). For example in the pojective plane (and due to 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 and so they have at least one real root.

Instead of discussing a general theory, we will illustrate now shortly its usefulness on several results for projective planes. .

Proposition. *A projective transformation different from the identity has either exactly one center and exactly one axis or it does not have center neither axis.*

Proof. Let us consider collineation f on $\mathcal{P}\mathbb{R}^3$ and let us assume that it has two distinct centers A and B. Let us denote by ℓ the line given by these two centers, and let us choose a point X in projective plane outside of ℓ. If p and q are the lines passing through pairs of points (A, X) respectively (B, X), then also $f(p) = p$ and $f(q) = q$ and in particular also point X is fixed. But this means that 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 means that f is identity mapping and so we proved that every projective transformation different from the identity may have at most one center. The same consideration for the dual projective plane gives the result about at most one axis.

If f has a center A, then all lines passing through A are fixed and correspond therefore to a two–dimensional subspace of row eigenvectors of matrix corresponding to transformation f. Therefore, there exists a two–dimensional subspace of column eigenvectors to the same eigenvalue, and this one represents exactly the line of fixed points, hence 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. □

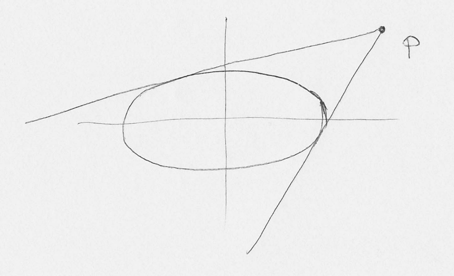
For practical problems it is useful to work with complex projective extensions also in the case of a real plane. Then the geometric behaviour can be easily read off the potential existence of real or imaginary centers and axes.

4.42. Projective classification of quadrics. In the end of this section we come back to conics and quadrics. A quadric Q in n–dimensional affine space $\mathbb{R}^n$ is defined by general quadratic equation (4.4), see page 228. Viewing affine space $\mathbb{R}^n$ as affine coordinates in projective space $\mathcal{P}\mathbb{R}^{n+1}$ we may aim to describe the set Q by homogeneous coordinates in projective space. The formula in these coordinates should contain only terms of second order since only a homogeneous formula is independent of the choice of the multiple of homogeneous coordinates $(x_0, x_1, \dots, x_n)$ of a point. Hence we are searching for a homogeneous formula whose restriction to affine coordinates, i.e. substitution $x_0 = 1$, gives the original formula (4.4).

But this is especially easy, we simply add enough x_0 to all terms – no one to quadratic terms, one to linear terms and x_0^2 to the constant term in the original affine equation for Q.

So we get a well defined quadratic form f on vector space $\mathbb{R}^{n+1}$ whose zero set defines correctly so called *projective quadric* $\bar{Q}$.

The intersection of "cone" $\tilde{Q} \subset \mathbb{R}^{n+1}$ of the zero set of this form with affine plane $x_0 = 1$ is the original quadric Q whose points are called proper points of the quadric, while the other points $\bar{Q} \setminus Q$ in the projective extension are the infinite points.



which means that every point X of line is lying on conic section f. However, when the conic section contains a line, it has to be degenerate, which is contradiction. As well, we can see that in the case of degenerate conic section, the polar is line of conic section itself.

Claim for $P \notin f$ follows from the corollary of symmetry of bilinear form F. When point Q lies on polar of P, then point P lies on polar of Q.

□

Using polar conjugates we can find axes and center of conic sections without need of Lagrange algorithm.

Consider 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 $\alpha = a_{33}$. That means the conic section is defined by equation

$$u^T \bar{A} u + 2a^T u + \alpha = 0$$

for vector $u = (x, y)$. Now we show that

4.58. Axes of conic section are polars of points at infinity determined by eigenvectors of matrix $\bar{A}$.

Solution. Because of symmetry of $\bar{A}$, in the basis of its eigenvectors, it has diagonal shape $D = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$, where $\lambda, \mu \in \mathbb{R}$ and this basis is ortogonal. Denote by U matrix changing basis to eigenbasis (columns are eigenvectors), then the conic section matrix in eigenbasis is

$$\begin{pmatrix} U^T & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{A} & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} D & U^T a \\ a^T U & \alpha \end{pmatrix}$$

So in this basis we have canonical form defined by vector $U^T a$ (up to translation). Specifically, denote by v_λ, v_μ eigenvectors and we have

$$\lambda(x + \frac{a^T v_\lambda}{\lambda})^2 + \mu(y + \frac{a^T v_\mu}{\mu})^2 = \frac{(a^T v_\lambda)^2}{\lambda} + \frac{(a^T v_\mu)^2}{\mu} - \alpha.$$

It means that eigenvectors are direction vectors of the conic section axes (so called main directions) and axes equations in this basis are $x = -\frac{a^T v_\lambda}{\lambda}$ and $y = -\frac{a^T v_\mu}{\mu}$. Axes coordinates u_λ and u_μ in 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 $v_\lambda^T(\lambda u_\lambda +$

The classification of real or complex projective quadrics, up to projective transformations, is a problem which we have already managed — it is all about finding the canonical polar basis, see paragraph 4.29. From this classification, which is given by signature of the form in real case and only by rank in complex case, we can deduce relatively easily also the classification of affine quadrics. We show the core of the procedure on the case of conics in affine and projective plane.

The projective classification gives the following possibilities, described by homogeneous coordinates $(x : y : z)$ in projective plane $\mathcal{P}\mathbb{R}^3$:

- imaginary regular conic given by $x^2 + y^2 + z^2 = 0$
- real regular conic given by $x^2 + y^2 - z^2 = 0$
- pair of imaginary lines given by $x^2 + y^2 = 0$
- pair of real lines given by $x^2 - y^2 = 0$
- one double line $x^2 = 0$.

We consider this classification as real, i.e. the classification of quadratic forms is given not only by its rank but also by its signature. Nevertheless, the points of quadric are considered also in the complex extension. In this way we should understand the stated names, e.g. the imaginary conic does not have any real points.

4.43. Affine classification of quadrics. For affine classification we must restrict the projective transformations to those which preserve the line of infinite points. We can realize this also by an opposite procedure — for a fixed projective type of conic Q, i.e. its cone $\tilde{Q} \subset \mathbb{R}^3$, we are choosing different affine planes $\alpha \subset \mathbb{R}^3$ which do not pass through origin, and we observe how the set of points $\tilde{Q} \cap \alpha$, which are proper points of Q in affine coordinates realized by plane α, is changing.

Hence in the case of a regular conic we have a true cone $\tilde{Q}$ given by equation $z^2 = x^2 + y^2$ and as planes α we may take the tangent planes to unite sphere for instance. If we begin with plane $z = 1$, the intersection consists only from finite points forming a unite circle Q. By a successive sloping of α we get more and more stretched ellipse until we get such slope that α is parallel with one of lines of the cone. In that moment there appears one (double) infinite point of our conic whose finite points still form one connected component, and so we get parabola. The continuation in sloping gives rise to two infinite points and the set of finite points is no more connected, and so we get the last regular quadric in the affine classification, a hyperbola.

We can take the advice from the introduced method which enable us to continue the classification in higher dimensions. In particular, let us notice that the intersection of our conic with the projective line of infinite points is always a quadric in dimension one less, i.e. in our case it is either an empty set or a double point or two points as types of quadrics on a projective line. Next we found out that 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 farther.

$a) = 0$ and $v_\mu^T(\mu u_\mu + a) = 0$. These equations are equivalent to equations $v_\lambda^T(\bar{A}u_\lambda + a) = 0$ a $v_\mu^T(\bar{A}u_\mu + a) = 0$ which are polar equations of points defined by vectors v_λ a v_μ. □

4.59. Remark. Corollary of the previous claim is a fact that center of the conic section is polar conjugated with all points at infinity. Center s coordinates then satisfy equation $\bar{A}s + a = 0$.

If $\det(A) \neq 0$, then equation $\bar{A}s + a = 0$ for center coordinates has exactly one solution for $\delta = \det(\bar{A}) \neq 0$ and no solutions for $\delta = 0$. That means that, regarding non-degenerate conic sections, ellipse and hyperbola have exactly one center and parabola does not have any (its center is point at infinity).

4.60. Prove that angle between tangent to parabola (with arbitrary touch point) and parabola axis is the same as angle between the tangent and line connecting focus and the touch point.

Solution. Polar (i.e. tangent) of point X=$[x_0, y_0]$ to parabola defined by canonical equation in polar basis is line satisfying

$$(x_0, y_0, 1)\begin{pmatrix}1 & 0 & 0\\ 0 & 0 & -p\\ 0 & -p & 0\end{pmatrix}\begin{pmatrix}x\\ y\\ 1\end{pmatrix} = x_0x - py - py_0 = 0$$

Cosine of angle between tangent and parabola axis ($x = 0$) is given by dot product of corresponding unit direction vectors. Unit direction vector of the tangent is $\frac{1}{\sqrt{p^2+x_0^2}}(p, x_0)$ and therefore for cosine we have

$$\frac{1}{\sqrt{p^2 + x_0^2}}(p, x_0).(0, 1) = \frac{x_0}{\sqrt{p^2 + x_0^2}}$$

Now we show that cosine of angle between tangent and line connecting focus F=$[0, \frac{p}{2}]$ and touch point X is equal. 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 cosine of angle we have

$$\frac{1}{\sqrt{p^2 + x_0^2}}\frac{1}{\sqrt{x_0^2 + (y_0 - \frac{p}{2})^2}}(x_0y_0 + \frac{px_0}{2})$$

Substituting $y_0 = \frac{x_0^2}{2p}$ we get $\frac{x_0}{\sqrt{p^2+x_0^2}}$.

This example shows that lightrays striking parallel with axis of parabolic mirrow are reflecting to the focus and, vice versa, lightrays going through focus reflect in direction parallel with axis of parabola. This is the principle of many devices such as parabolic reflector. □

4.61. Find equation of tangent in P=[1, 1] to the conic section

$$4x^2 + 5y^2 - 8xy + 2y - 3 = 0$$

Solution. By projecting we get conic section defined by 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 previous theorem, tanget is 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 we get tangent equation $y = 1$.

□

4.62. Find coordinates of intersection of y axis and conic section defined by

$$5x^2 + 2xy + y^2 - 8x = 0$$

Solution. y axis, i.e. line $x = 0$, is polar of sought point P with homogeneous coordinates $\langle p \rangle = (p_1 : p_2 : p_3)$. That meansthat equation $x = 0$ is equivalent to 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 us for 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 point P coordinates by 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 y axis touches the conic section in the origin. □

4.63. Find a touch point of line $x = 2$ with conic section from the previous exercise.

Solution. Line has equation $x - 2z = 0$ in projective extension and therefore we get condition $Ap = (\alpha, 0, -2\alpha)$ for touch point P, which gives us

$$\begin{aligned} 5p_1 + p_2 - 4p_3 &= \alpha \\ p_1 + p_2 &= 0 \\ -4p_1 &= -2\alpha \end{aligned}$$

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 $[2, -2]$. □

4.64. Find equations of tangents passing through P= $[3, 4]$ to tangent defined by

$$2x^2 - 4xy + y^2 - 2x + 6y - 3 = 0$$

Solution. Suppose that the touch point has homogeneous coordinates given by multiple of vector $t = (t_1, t_2, t_3)$. Condition of T lying on conic section is $t^T At = 0$, which gives

$$2t_1^2 - 4t_1t_2 + t_2^2 - 2t_1t_3 + 6t_2t_3 - 3t_3^2 = 0$$

Condition of P lying on polar of T is $p^T At = 0$, where $p = (3, 4, 1)$ are homogeneous coordinates of point P. In this case, the equation gives us

$$(3, 4, 1)\begin{pmatrix} 2 & -2 & -1 \\ -2 & 1 & 3 \\ -1 & 3 & -3 \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = -3t_1 + t_2 + 6t_3 = 0$$

Now we can input $t_2 = 3t_1 - 6t_3$ to the previous quadratic equation. Then

$$-t_1^2 + 4t_1t_3 - 3t_3^2 = 0$$

Because for $t_3 = 0$ equation is not satisfied, we can 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 $(3 : 3 : 1)$. Tangent equations are polars of those points $7x - 2y - 13 = 0$ and $x = -3$. □

4.65. Find equation of tangent passing through origin to the circle

$$x^2 + y^2 - 10x - 4y + 25 = 0$$

Solution. 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 derive for example t_2 and substitute in circle equation, which $(t_1 : t_2 : t_3)$ has to satisfy as well. We get quadratic equation $29t_1^2 - 250t_1 + 525 = 0$, which has solutions $t_1 = 5$ and $t_1 = \frac{105}{29}$. We compute 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 $y = 0$ a $20x - 21y = 0$. □

4.66. Find tangents equations to circle $x^2 + y^2 = 5$ which are parallel with $2x + y + 2 = 0$.

Solution. In projective extension, these tangets intersect in 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. 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 satisfy $t_1 - 2t_2 = 0$. Substituting into circle equation we get $5t_2^2 = 5$. Since that $t_2 = \pm 1$ and touch points are $[2, 1]$ and $[-2, -1]$. □

Tangent touching the conic section at infinity is called conic section *asymptote*. Number of asymptotes of conic section is equal to number of intersections between conic section and line at infinity, which means that ellipse does not have any real asymptote, parabola has one (which is however line at infinity) and hyperbola two of them.

4.67. Find points at infinity and asymptotes of conic section defined by

$$4x^2 - 8xy + 3y^2 - 2y - 5 = 0$$

Solution. First, we rewrite the conic section in homogeneous coordinates.

$$4x^2 - 8xy + 3y^2 - 2yz - 5z^2 = 0$$

Points at infinity are then points determined by homogeneous coordinates $(x : y : 0)$ satisfying this equation, which means

$$4x^2 - 8xy + 3y^2 = 0.$$

For fraction $\frac{x}{y}$ we get two solutions: $\frac{x}{y} = -\frac{1}{2}$ and $\frac{x}{y} = -\frac{3}{2}$. Conic section is therefore hyperbola with points at infinity P= $(-1 : 2 : 0)$ a Q= $(-3 : 2 : 0)$. Asymptoty jsou potom polĂĄry bodĹŻ P a Q, tj.

$$(-1, 2, 0)\begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -12x + 10y - 2 = 0$$

a

$$(-3, 2, 0)\begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -20x + 18y - 2 = 0$$

□

You can find further exercises on conic sections on the page 250.

4.68. **Harmonic cross-ratio.** If cross-ratio of four points lying on line equal to -1, we talk about *harmonic quadruple*. Let $ABCD$ be a quadrilateral. Denote by K intersection of lines AB and CD, by M intersection of lines AD and BC. Further let L, N be intersection of KM and AC, BD respectively. Show that points K, L, M, N are harmonic quadruple. ○

D. Further exercise on this chapter

4.69. Find parametric equation of intersection of planes in $\mathbb{R}^3$:

$$\sigma : 2x + 3y - z + 1 = 0 \quad \text{a} \quad \rho : x - 2y + 5 = 0.$$

○

4.70. Find common perpendicular of skew lines

$$p : [1, 1, 1] + t(2, 1, 0), \quad q : [2, 2, 0] + t(1, 1, 1).$$

○

4.71. Jarda is standing in $[-1, 1, 0]$ and has a stick of length 4. Can he simultaneously touch lines p and q, where

$$\begin{aligned} p &: [0, -1, 0] + t(1, 2, 1), \\ q &: [3, 4, 8] + s(2, 1, 3)? \end{aligned}$$

○ (Stick has to pass through $[-1, 1, 0]$.)

4.72. Cube $ABCDEFGH$ is given. Let point T lie on edge BF, $|BT| = \frac{1}{4}|BF|$. Compute cosine of angle between ATC and BDE. ○

4.73. Cube $ABCDEFGH$ is given. Let point T lie on edge AE, $|AT| = \frac{1}{4}|AE|$ and S is midpoint of AD. Compute cosine of angle between BDT and SCH. ○

4.74. Cube $ABCDEFGH$ is given. Let point T lie on edge BF, $|BT| = \frac{1}{3}|BF|$. Compute cosine of angle between ATC and BDE. ○

4.75. Determine tangent to ellipse $\frac{x^2}{16} + \frac{y^2}{9} = 1$ parallel with line $x + y - 7 = 0$.

Solution. Lines parallel with given line intersect this line in point at infinity $(1 : -1 : 0)$. We construct tangents to given ellipse passing through this point. Touch point 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 in ellipse equation we get $t_1 = \pm\frac{16}{5}$. Touch points of sought tangents are $[\frac{16}{5}, \frac{9}{5}]$ and $[-\frac{16}{5}, -\frac{9}{5}]$. Tangents are polars of those points. These have equations $x + y = 5$ and $x + y = -5$. □

4.76. Determine points at infinity and asymptotes of conic section

$$2x^2 + 4xy + 2y^2 - y + 1 = 0$$

Solution. Equation of points at infinity $2x^2 + 4xy + 2y^2 = 0$ or rather $2(x + y)^2 = 0$ has solution $\frac{x}{=} - y$. The only point at infinity therefore is $(1 : -1 : 0)$ (conic section is a parabola). Asymptote is a polar of this point, specifically line at infinity $z = 0$. □

4.77. Prove that product of distances between arbitrary point on a hyperbola and its asymptotes is consant and tell the value of this constant.

Solution. Denote the point lying on hyperbola as P. Asymptote equation of hyperbola in canonical form is $bx \pm ay = 0$. Their normals are $(b, \pm a)$ and from here we determine projections P_1, P_2 of point P to asymptotes. For 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 $\frac{a^2q^2 - b^2p^2}{a^2+b^2} = \frac{a^2b^2}{a^2+b^2}$, because P lies on hyperbola. □

4.78. Compute angle between asymptotes of hyperbola $3x^2 - y^2 = 3$.

Solution. For cosine of angle between asymptotes of hyperbola in canonical form we get $\cos\alpha = \frac{b^2-a^2}{b^2+a^2}$. In our case the angle is equal 60°. □

4.79. Compute centers of 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) System $\bar{A}s + a = 0$ for computing proper centers is

$$\begin{aligned} 9s_1 + 3s_2 &= 0 \\ 3s_1 - 2 &= 0 \end{aligned}$$

and, solving it, we obtain center $[\frac{2}{3}, -2]$.

(b) In this case we have

$$\begin{aligned} s_1 + s_2 + 1 &= 0 \\ s_1 + s_2 + \tfrac{1}{2} &= 0 \end{aligned}$$

and therefore there is no proper center (conic section is a parabola). Moving to homogeneous coordinates we can obtain center at infinity $(1 : -1 : 0)$.

(c) Coordinates of center in this case satisfy

$$\begin{aligned} s_1 - 2s_2 + 1 &= 0 \\ -2s_1 + 4s_2 - 2 &= 0 \end{aligned}$$

and the solution is whole line of centers. It is so because the conic section is degenerated to pair of parallel lines.

(d) From equations for center computation we immediately get that center is (α, β). Coordinates of center therefore gives translation of coordinate system origin to the frame in which the ellipse has basic form.

□

4.80. Tell the equations of axes of conic section $6xy + 8y^2 + 4y + 2x - 13 = 0$.

Solution. Main directions of the conic section (axes direction vectors) are eigenvectors of matrix $\begin{pmatrix} 0 & 3 \\ 3 & 8 \end{pmatrix}$. Characteristic equation has form $\lambda^2 - 8\lambda - 9 = 0$ and eigenvalues are therefore $\lambda_1 = -1$, $\lambda_2 = 9$. Corresponding eigenvectors are then $(3, -1)$ and $(1, -3)$. Axes are polars of points at infinity defined by those directions. For $(3, -1)$ we get axis equation $-3x + y + 1 = 0$ and for $(1, -3)$ axis $-9x - 21y - 5 = 0$. □

4.81. Determine equations of axes of conic section $4x^2 + 4xy + y^2 + 2x + 6y + 5 = 0$.

Solution. Eigenvalues of matrix $\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$ are $\lambda_1 = 0$, $\lambda_2 = 5$ and corresponding eigenvectors are $(-1, 2)$ and $(2, 1)$. We get axes $5 = 0$ and $2x + y + 1 = 0$. The former axis is obviously satisfied for no point. Hence there is only on axis (the conic section is a parabola). □

4.82. Equation

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

defines a conic section. Tell its center, axes, asymptotes and foci.

Exercise solution

4.9. 2, 3, 4, 6, 7, 8. Try to find planes positions which correspond to each of those numbers on your own.

4.28. For normal vector (a, b, c) of such planes we have $a+b = 0$ (ortogonal to p) 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 $x - y \pm \sqrt{6} - 1 = 0$.

4.33. $(-1, 3, 2)$.

4.69. Line $(2t, t, 7t) + [-5, 0, -9]$.

4.70. $[3, 2, 1][8/3, 8/3, 2/3]$.

4.71. Transversal $[1, 1, 1][-3, 1, -1]$ of length $\sqrt{20}$, stick is not long enough.

4.72. $\frac{2\sqrt{6}}{9}$

4.73. $\frac{\sqrt{3}}{6}$.

4.74. $\frac{\sqrt{3}}{\sqrt{11}}$

CHAPTER 5

Establishing the ZOO

which functions do we need for our models?
– a thorough menagerie

A. Polynomial interpolation

At the beginning of this chapter, 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 as well. We will try to find a polynomial (of the least degree possible) satisfying these dependencies.

5.1. 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 $a_3x^3 + a_2x^2 + a_1x_1 + a_0$. We know there is exactly one polynomial of degree less than four and satisfying the given conditions.

$$\begin{aligned} a_0 + 2a_1 + 4a_2 + 8a_3 &= 1 \\ a_0 + 3a_1 + 9a_2 + 27a_3 &= 0 \\ a_0 + 4a_1 + 16a_2 + 64a_3 &= -1 \\ a_0 + 5a_1 + 25a_2 + 125a_3 &= 6. \end{aligned}$$

In this chapter, we begin to develop tools that will allow us to model dependencies which are neither linear, nor discrete. We can often meet this need when we have a system developing in time and we try to describe it not only at several moments, but "continuously", i. e. for all possible points in time. Sometimes this is an intent (this concerns, for instance, physical models of classical mechanics), whereas other times it may be an appropriate approach to discrete models (in economics, chemistry, or biology, for example).

The key concept is that of a function. The larger class of functions we admit, the more difficult it will be to develop the necessary tools for our work. On the other hand, if there are only few types of functions available, it may happen that we will not be able to model some real situations at all. The objective of the following two chapters is thus to explicitly introduce several types of elementary functions, to implicitly describe far more functions, and to build the standard tools for the work with them. This is called differential and integral calculus of one variable. While, so far, we have mainly focused on the part of mathematics called *algebra*, now we will approach the so-called *mathematical analysis*.

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

Let us remind the discussion from the paragraph 1.4, where we thought about how to deal with scalar functions. There is nothing to add to this discussion and we would like (to start off) to work with functions $\mathbb{R} \to \mathbb{R}$ (*real-valued functions of a real variable*), or $\mathbb{R} \to \mathbb{C}$ (*complex-valued functions of a real variable*), or functions $\mathbb{Q} \to \mathbb{Q}$ (rational-valued functions of a rational variable) and so on. Our conclusions can usually be extended to the cases concerning vector values over the considered scalars, but we will mostly talk only about the cases of real and complex numbers.

Let us begin with the easiest functions which we can assign explicitly by finitely many algebraic operations with scalars.

5.1. Polynomials. We can add and multiply scalars, and these operations satisfy a number of properties which we enumerated in the paragraphs 1.1 and 1.3. If we admit any finite number of these operations, leaving one of the variables as an unknown and fixing the other scalars, we get the so-called polynomial functions:

Each equation arose from one of the given conditions.

Another option is to construct the required polynomial from the fundamental Lagrange polynomials.

(see 5.4):

$$\begin{aligned}P(x) &= 1\cdot\frac{(x-3)(x-4)(x-5)}{(2-3)(2-4)(2-5)}+0\cdot(\dots)+\\ &= (-1)\cdot\frac{(x-2)(x-3)(x-5)}{(4-2)(4-3)(4-5)}+6\cdot\frac{(x-2)(x-3)(x-4)}{(5-2)(5-3)(5-4)}=\\ &= \frac{4}{3}z^3-12z^2+\frac{101}{3}z-29.\end{aligned}$$

The coefficients of the polynomial form, of course, the solution of the aforementioned system of linear equations. □

5.2. Find a polynomial P satisfying the following conditions:

$$P(1+i)=i,\ P(2)=1,\ P(3)=-i.$$

5.3. For pairwise distinct points $x_0,\dots,x_n\in\mathbb{R}$, consider the elementary Lagrange polynomials (5.4)

$$l_i(x):=\frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)},\quad x\in\mathbb{R},\ i=0,\dots,n.$$

Prove that

$$\sum_{i=0}^{n}l_i(x)=1\quad\text{for all } x\in\mathbb{R}.$$

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 $x_0,\dots,x_n$. However, there is exactly one such polynomial, namely the constant polynomial $y\equiv 1$. □

5.4. 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

Polynomials

A *polynomial* over a ring of scalars $\mathbb{K}$ is a mapping $f:\mathbb{K}\to\mathbb{K}$ given by the expression

$$f(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0,$$

where a_i, $i=0,\dots,n$, are fixed scalars, multiplication is indicated by mere concatenation of symbols, and "+" denotes addition. If $a_n\neq 0$, we say that the the polynomial f has *degree n*. The degree of the zero polynomial is undefined. The scalars a_i are called *coefficients of the polynomial f*.

The polynomials of degree zero are exactly the non-zero constant mappings. In algebra, polynomials are more often defined as formal expressions of the aforementioned form of $f(x)$, i. e. a polynomial is defined to be a sequence $a_0,a_1,\dots$ of coefficients such that only finitely many of them are non-zero. However, we will show shortly that these approaches are equivalent.

It is easy to verify that the polynomials over a given ring of scalars form a ring as well. The multiplication and addition of polynomials are given by the operations in the original ring $\mathbb{K}$ by the values of the polynomials, i. e.

$$(f\cdot g)(x)=f(x)\cdot g(x),\quad (f+g)(x)=f(x)+g(x),$$

where the operations on the left-hand side are interpreted in the ring of polynomials whereas the operations on the right-hand side are the ones of the ring of scalars.

5.2. Euclidean division of polynomials. As we have already mentioned, we will work exclusively with the scalar fields $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$. In all these fields, the following statement holds:

Proposition (Euclidean division of polynomials)**.** *For any two polynomials f of degree n and g of degree m, there is exactly one pair of polynomials q, r such that $f=q\cdot g+r$ and the degree of the polynomial r is less than m or $r=0$.*

Proof. Let us begin with the uniqueness. Suppose we have two expressions of the polynomial f in terms of the polynomials q, q', r, and r', i. e. we have

$$f=q\cdot g+r=q'\cdot g+r'.$$

Subtraction gives $0=(q-q')\cdot g+(r-r')$.

If $q=q'$, then $r=r'$ as well. And if $q\neq q'$, then the term of the highest degree in $(q-q')\cdot g$ cannot be compensated by $r-r'$, which leads to a contradiction. We have thus proved the uniqueness of the result of the division, supposing it exists.

It remains to prove that the polynomial f can always be expressed in the wanted form. If $m>n$, we can immediately set $f=0\cdot g+f$. Therefore, let us suppose that $n\geq m$ and prove the proposition by induction on the degree of the polynomial f.

If f is of degree zero, then the statement is trivial. Let us thus suppose that the statement holds for all polynomials f of degree less than $n>0$ and consider the expression $h(x)=f(x)-\frac{a_n}{b_m}x^{n-m}g(x)$. Either $h(x)$ is the zero polynomial and we have got what we have been looking for, or it is a polynomial of a lower degree and as such can be written in the desired form as $h(x)=q\cdot g+r$ whence

$$f(x)=h(x)+\frac{a_n}{b_m}x^{n-m}g(x)=(q+\frac{a_n}{b_m}x^{n-m})g(x)+r$$

and we are done. □

of degree less than four, we get the same number of equations and unknown coefficients (let us say $P(x) = a_3x^3 + a_2x^2 + a_1x + a_0$):

$$\begin{aligned} P(1) = \quad & (a_3 + a_2 + a_1 + a_0 & = 0, \\ P'(1) = \quad & 3a_3 + 2a_2 + a_1 = & 1, \\ P(2) = \quad & 8a_3 + 4a_2 + 2a_1 + a_0 = & 3, \\ P'(2) = \quad & 12a_3 + 4a_2 + a_1 = & 3. \end{aligned}$$

By solving this system, we obtain the polynomial $P(x) = -2x^3 + 10x^2 - 13x + 5$.

Another solution. We will use fundamental Hermite polynomials:

$$\begin{aligned} h_1^1(x) &= \left(1 - \frac{2}{0 + (-1)}(x-1)\right)(2-x)^2 = (2x-1)(x-2)^2, \\ h_2^1(x) &= (5-2x)(x-1)^2, \\ h_1^2(x) &= (x-1)(x-2)^2, \\ h_2^2(x) &= (x-2)(x-1)^2. \end{aligned}$$

Altogether,

$$P(x) = 0 \cdot h_1^1(x) + 3 \cdot h_2^1(x) + 1 \cdot h_1^2(x) + 3 \cdot h_2^2(x) = -2x^3 + 10x^2 - 13x + 5.$$

□

5.5. 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,

$$\begin{aligned} P(1) &= \frac{1}{2} \cdot 8\frac{(\pi-3)(\pi-2)}{\pi^2} - \frac{1}{4} \cdot 9\frac{(\pi-4)(\pi-2)}{\pi^2} + 0 = \\ &= \frac{(5\pi-12)(\pi-2)}{4\pi^2} \doteq 0.288913. \end{aligned}$$

We may notice we did not need to calculate the third elementary polynomial. The actual value is $\cos^2 1 \doteq 0.291927$. □

If the value $f(b)$ equals zero for some element $b \in \mathbb{K}$, then we must have $r = 0$ in the quotient $f(x) = q(x)(x-b) + r$ because otherwise we could not achieve $f(b) = q(b) \cdot 0 + r$, where the degree of r is zero. We say that b is a *root of the polynomial* f. The degree of q is then exactly $n-1$. If q also has a root, we can continue and in no more than n steps we arrive at a constant polynomial. Thus we have proved that the number of roots of any non-zero polynomial over the field $\mathbb{K}$ is at most the degree of the polynomial. Hence we can easily derive the following observation:

Corollary. *If $\mathbb{K}$ is an infinite field, then the polynomials f and g are equal as mappings if and only if they equal as sequences of coefficients.*

Proof. Suppose that $f = g$, i. e. $f - g = 0$, as a mapping. Therefore, the polynomial $(f-g)(x)$ has infinitely many roots, which is possible only if it is the zero polynomial. □

Let us realize that of course, this statement does not hold with finite fields. A simple non-example is the polynomial $x^2 + x$ over $\mathbb{Z}_2$ which represents a constant zero mapping.

5.3. Interpolation polynomial. It is often useful to give an easily computable expression for a function which is given by the values it takes at some given points $x_0, \dots, x_n$. If the values were all zeros, we can immediately find a polynomial of degree $n+1$, namely

$$f(x) = (x - x_0)(x - x_1) \dots (x - x_n),$$

which takes zero at these points and only at them. However, there are more polynomials which takes zero at the given points, for 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 $x_0, \dots, x_n \in \mathbb{K}$ and the given values $y_0, \dots, y_n \in \mathbb{K}$ is the zero polynomial or a polynomial of degree at most n such that $f(x_i) = y_i$ for all $i = 0, 1, \dots, n$.

Theorem. *For every set of $n+1$ (pairwise distinct) points $x_0, \dots, x_n \in \mathbb{K}$ and the given values $y_0, \dots, y_n \in \mathbb{K}$, there is exactly one interpolation polynomial f.*

Proof. Let us begin with the easier part, i. e. the uniqueness. 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 $f - g = 0$.

It remains to prove the existence. Let us label the coefficients of the polynomial f of degree n:

$$f = a_nx^n + \dots + a_1x + a_0.$$

5.6. 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:

$$\begin{aligned}
l_0(x) &= \frac{(x-\frac{\pi}{6})(x-\frac{\pi}{4})(x-\frac{\pi}{3})(x-\frac{\pi}{2})}{(0-\frac{\pi}{6})(0-\frac{\pi}{4})(0-\frac{\pi}{3})(0-\frac{\pi}{2})} \\
&\doteq 1.4783x^4 - 5.8052x^3 + 8.1057x^2 - 4.7746x + 1, \\
l_1(x) &= \frac{(x-0)(x-\frac{\pi}{4})(x-\frac{\pi}{3})(x-\frac{\pi}{2})}{(\frac{\pi}{6}-0)(\frac{\pi}{6}-\frac{\pi}{4})(\frac{\pi}{6}-\frac{\pi}{3})(\frac{\pi}{6}-\frac{\pi}{2})} \\
&\doteq -13.3046x^4 + 45.2808x^3 - 49.2419x^2 + 17.1887x, \\
l_2(x) &= \frac{(x-0)(x-\frac{\pi}{6})(x-\frac{\pi}{3})(x-\frac{\pi}{2})}{(\frac{\pi}{4}-0)(\frac{\pi}{4}-\frac{\pi}{6})(\frac{\pi}{4}-\frac{\pi}{3})(\frac{\pi}{4}-\frac{\pi}{2})} \\
&\doteq 23.6526x^4 - 74.3070x^3 + 71.3298x^2 - 20.3718x, \\
l_3(x) &= \frac{(x-0)(x-\frac{\pi}{6})(x-\frac{\pi}{4})(x-\frac{\pi}{2})}{(\frac{\pi}{3}-0)(\frac{\pi}{3}-\frac{\pi}{6})(\frac{\pi}{3}-\frac{\pi}{4})(\frac{\pi}{3}-\frac{\pi}{2})} \\
&\doteq -13.3046x^4 + 38.3146x^3 - 32.8279x^2 + 8.5943x, \\
l_4(x) &= \frac{(x-0)(x-\frac{\pi}{6})(x-\frac{\pi}{4})(x-\frac{\pi}{3})}{(\frac{\pi}{2}-0)(\frac{\pi}{2}-\frac{\pi}{6})(\frac{\pi}{2}-\frac{\pi}{4})(\frac{\pi}{2}-\frac{\pi}{3})} \\
&\doteq 1.4783x^4 - 3.4831x^3 + 2.6343x^2 - 0.6366x.
\end{aligned}$$

Then, the value of the interpolation polynomial is

$$P(x) = 0 \cdot {}_0(x) + \frac{1}{2}l_1(x) + \frac{\sqrt{2}}{2}l_2(x) + \frac{\sqrt{3}}{2}l_3(x) + l_4(x) \doteq$$

$$\doteq 0.0288x^4 - 0.2043x^3 + 0.0214x^2 + 0.9956x.$$

□

Additional questions: Can Joe use this formula to calculate the sine's values at the interval $[\frac{\pi}{2}, \pi]$? 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?

5.7. 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

Substituting the wanted values leads to a system of $n+1$ equations for the same number of unknown coefficients a_i

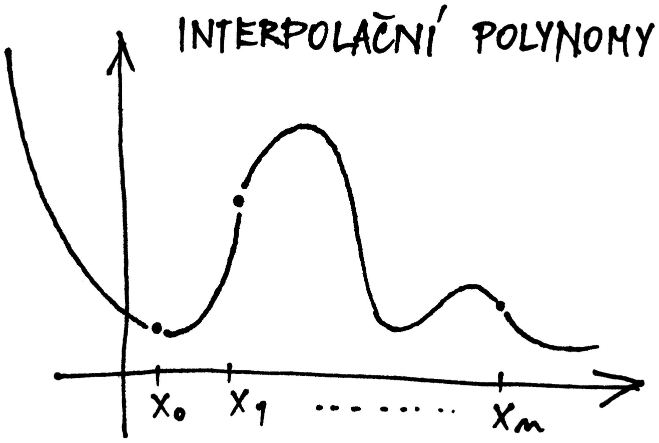
$$a_0 + x_0a_1 + \cdots + (x_0)^n a_n = y_0$$
$$\vdots$$
$$a_0 + x_na_1 + \cdots + (x_n)^n a_n = y_n.$$

The existence of a solution of this system can easily be shown by straight construction of the polynomial by the so-called Lagrange polynomials for the given points $x_0, \dots, x_n$, see the next paragraph.

However, we will now finish the proof using only our basic knowledge from linear algebra. This system of linear equations has a unique solution if the determinant of its matrix is an invertible scalar, i. e. a non-zero scalar (see 3.1 and 2.23). It is the so-called *Vandermonde determinant*, which was discussed in the exercise ||2.24|| on page 87.

Since we have verified that for zero right-hand sides, there is exactly one solution, we know that this determinant must be non-zero.

And since polynomials equal as mappings iff they equal as sequences of coefficients, the theorem is proved. □



5.4. Applications of interpolations. At first sight, it may seem that real or rational polynomials, i. e. polynomial functions $\mathbb{R} \to \mathbb{R}$ or $\mathbb{Q} \to \mathbb{Q}$, form a very nice class of functions of one variable. We can lay them through any set of given values. Moreover, they are easily expressible, so their value at any point can be calculated without difficulties. However, we encounter a number of problems when trying to put them in practice.

The first of the problems is to quickly find the polynomial which we will lay through the given data since solving the aforementioned system of linear equations generally requires time proportional to the cubed number of given points, which is unacceptable for larger data. Another problem is slow computation of the value of a polynomial of a relatively high degree at a given point. Both of these problems can be partially bypassed by selecting a more convenient expression of the interpolation polynomial (i. e. we choose a basis of the corresponding vector space of all polynomials of degree at most k which is better than the standard basis $1, x, x^2, \dots, x^n$).

We will demonstrate this on one exercise:

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. □

5.8. 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 their roles, thus obtaining the elementary Lagrange polynomials

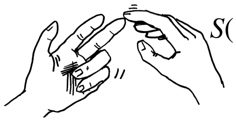
$$\begin{aligned} l_0(x) &= \frac{(x+0.75)(x-2)}{(-3+0.75)(-3-2)} = \frac{4}{45}x^2 - \frac{1}{9}x - \frac{2}{15}, \\ l_1(x) &= -\frac{16}{99}x^2 - \frac{16}{99}x + \frac{32}{33}, \\ l_2(x) &= \frac{6}{55}x^2 + \frac{3}{11}x + \frac{9}{55}. \end{aligned}$$

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? □

5.9. 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 and S_2 be zero at the points -1 and 1, respectively. Thanks to the given value at 0, we know that the absolute coefficients of both the polynomials are 1. By symmetry, the common value of the first derivative at the point 0 is 0. So we can set $S_1(x) = ax^3 + bx^2 + 1$ and $S_2(x) = cx^3 +$

Lagrange interpolation polynomials

Lagrange interpolation polynomial can easily be expressed in terms of the so-called elementary Lagrange polynomials ℓ_i of degree n with the properties

$$\ell_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Apparently, these polynomials must be (up to a constant) equal to the expressions $(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)$, and so

$$\ell_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

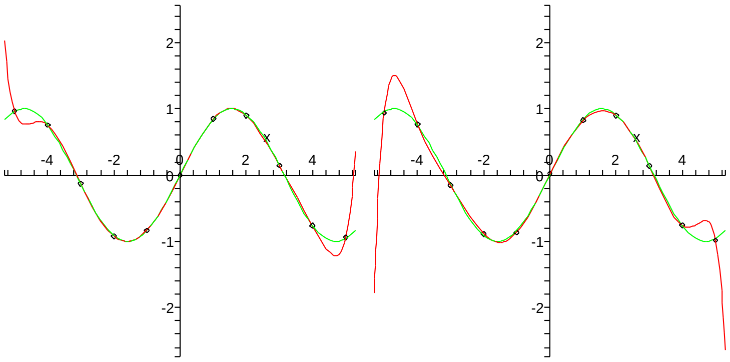
The wanted Lagrange interpolation polynomial is then given by

$$f(x) = y_0\ell_0(x) + y_1\ell_1(x) + \dots + y_n\ell_n(x).$$

The usage of Lagrange polynomials is especially efficient if we are working with different values y_i for the same set of values x_i because in this case, the elementary polynomials ℓ_i are already prepared.

One of the disadvantages of this expression is great sensitivity to inaccuracies in calculation when the differences of the given values x_i are small since one divides by these differences.

Another disadvantage is miserable stability of the values of real or rational polynomials as the variable grows. We will soon develop tools for an exact description of the functions' behavior, but even without them, it is clear that according to the sign of the coefficient of the term with highest degree, the polynomial's value will rapidly approach plus or minus infinity as x increases. However, the mentioned sign is even not stable under small changes of the defining values. This is illustrated by the following two pictures, displaying eleven values of the function $\sin(x)$ with small random changes of the values. There is the approximated function, the circles are gently moved values and the uniquely determined interpolation polynomial. While the approximation is quite good inside the interval, it is tremendous at the margins.



There is a rich theory about the interpolation polynomials, interested readers are advised to look at the special literature.

5.5. Note. The numerical instability caused by the closeness of (some) of the points x_i is clearly seen on the system of equations from the proof of the Theorem 5.3. When solving a system of linear equations, the instability is closely related to the magnitude of the

dx^2+1, where a, b, c, d are unknown real parameters. Confronting these forms with the conditions $S_1(-1)=0$, $S_1"(-1)=0$, $S_2(1)=0$, and $S_2"(1)=0$ yields the following system of four linear equations in the mentioned parameters.

$$\begin{aligned} -a+b+1 &= 0,\\ -6a+2b &= 0,\\ c+d+1 &= 0,\\ 6c+2d &= 0. \end{aligned}$$

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,

$$S(x)=\begin{cases} -\frac{1}{2}x^3-\frac{3}{2}x^2+1 & \text{pro } x\in[-1,0],\\ \frac{1}{2}x^3-\frac{3}{2}x^2+1 & \text{pro } x\in[0,1]. \end{cases}$$

□

5.10. Find a natural spline S which satisfies

$$S(-1)=0,\ S(0)=1,\ S(1)=0,\ S'(-1)=1,\ S'(1)=1.$$

Solution. Our spline differs from the previous one 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 $S_2(x)=cx^3+dx^2+1$, 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

$$\begin{aligned} -a+b+1 &= 0,\\ 3a-2b &= 1,\\ c+d+1 &= 0,\\ 3c+2d &= 1, \end{aligned}$$

having the solution $a=-1$, $b=-2$, $c=3$, $d=-4$. Hence, the wanted spline is the function

$$S(x)=\begin{cases} -x^3-2x^2+1 & \text{pro } x\in[-1,0],\\ 3x^3-4x^2+1 & \text{pro } x\in[0,1]. \end{cases}$$

□

5.11. 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. ○

5.12. Construct the Lagrange interpolation polynomial for

determinant of the corresponding matrix, i. e. the Vandermonde determinant, in our case. This can easily be calculated:

Lemma. *For any sequence of pairwise distinct scalars $x_0,\dots,x_n\in\mathbb{K}$, it holds that*

$$V(x_0,\dots,x_n)=\prod_{i>k=0}^{n}(x_i-x_k).$$

Proof. We will proof this statement by induction on the number of the points x_i. Apparently, it holds for $n=1$ (and the problem is completely uninteresting for $n=0$). Let us suppose that the result is correct for $n-1$, i. e.

$$V(x_0,\dots,x_{n-1})=\prod_{i>k=0}^{n-1}(x_i-x_k).$$

Now let us consider the values $x_0,\dots,x_{n-1}$ to be fixed and let us vary the value of x_n. Expanding the determinant by the last row (see **??**), we obtain the wanted determinant as the polynomial

$$(5.1)\qquad V(x_0,\dots,x_n)=(x_n)^nV(x_0,\dots,x_{n-1})-(x_n)^{n-1}\cdots.$$

This is a polynomial of degree n since we know that its coefficient at $(x_n)^n$ is non-zero, by the induction hypothesis. Apparently, it will take zero at any point $x_n=x_i$ for $i<n$ because in that case, the original determinant contains two identical rows. Our 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. Hence it follows that the whole Vandermonde determinant (as a polynomial in the variable x_n) must, up to a multiplicative constant, equal this expression, i. e.

$$V(x_0,\dots,x_n)=c\cdot(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1}).$$

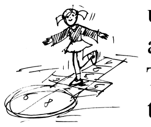
Confronting the coefficients at the highest exponent in (5.1) with this expression yields

$$c=V(x_0,\dots,x_{n-1}),$$

which finishes the proof of this lemma. □

Again, we can see that the determinant will be very small if the distances of the points x_i are such.

5.6. Derivatives of polynomials. We have found out that the values of the polynomials rapidly tend to infinite values as the input variable grows (see the pictures as well). Therefore, it is apparent that 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 $x\in\mathbb{R}$ is well expressed by

$$(5.2)\qquad \frac{f(x+\Delta x)-f(x)}{\Delta x},$$

and since we can calculate (over an arbitrary ring)

$$(x+\Delta x)^k=x^k+kx^{k-1}\Delta x+\dots+\binom{k}{l}x^l(\Delta x)^{k-l}+\dots+(\Delta x)^k,$$

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. ○

5.13. Find a polynomial $p(x) = ax^3 + bx^2 + cx + d$ which satisfies

$$p(0) = 1, \quad p(1) = 0, \quad p(2) = 1, \quad p(3) = 10.$$

○

5.14. Construct a polynomial p of degree three or less which satisfies

$$p(0) = 2, \quad p(1) = 3, \quad p(2) = 12, \quad p(5) = 147.$$

○

5.15. Let the values $y_0, \dots, y_n \in \mathbb{R}$ at pairwise distinct points $x_0, \dots, x_n \in \mathbb{R}$, respectively, be given. How many polynomials of degree exactly $n + 1$ and taking the given values at the given points are there? ○

5.16. Determine the Hermite interpolation polynomials P, Q satisfying

$$P(-1) = -11, \quad P(1) = 1, \quad P'(-1) = 12, \quad P'(1) = 4;$$
$$Q(-1) = -9, \quad Q(1) = -1, \quad Q'(-1) = 10, \quad Q'(1) = 2.$$

○

5.17. Replace the function f with a Hermite polynomial, knowing that

x_i	-1	1	2
$f(x_i)$	4	-4	-8
$f'(x_i)$	8	-8	11

.

○

5.18. Without calculation, determine the Hermite interpolation polynomial if the following is given:

$$x_0 = 0, \quad x_1 = 2, \quad x_2 = 1,$$
$$y_0 = 0, \quad y_1 = 4, \quad y_2 = 1,$$
$$y'_0 = 0, \quad y'_1 = 4, \quad y'_2 = 2.$$

○

5.19. 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. ○

we get, for the polynomial $f(x) = a_n x^n + \dots + a_0$, the above quotient in the form

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} = a_n\frac{nx^{n-1}\,\Delta x+\dots+(\Delta x)^k}{\Delta x} + \dots + a_1\frac{\Delta x}{\Delta x}$$
$$= na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + a_1 + \Delta x(\dots)$$

where the expression in parentheses is polynomially dependent on Δx. Clearly, for the values Δx very close to zero, we get a value arbitrarily close to the following expression:

Derivatives of polynomials

The derivative of a polynomial $f(x) = a_n x^n + \dots + a_0$ with respect to the variable x is the polynomial

$$f'(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + a_1.$$

>From the definition, it is clear that it is just the value $f'(x_0)$ of the derivative which gives us a good approximation of the polynomial's behavior near the point x_0. To be more precise, the lines

$$y = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}(x - x_0) + f(x_0),$$

i. e. 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 must be the tangent to the graph of the polynomial f. We talk about *linear approximation* of the polynomial f by its *tangent line*.

The derivative of polynomials is a linear mapping which assigns to polynomials of degree at most n polynomials of degree at most $n - 1$.

Iterating this procedure, we obtain 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 the so-called cyclic nilpotent mappings, which are more thoroughly examined in the paragraph 3.32 on nilpotent mappings.

5.7. Hermite's interpolation problem. Again, let us consider $m+1$ pairwise distinct real numbers $x_0, \dots, x_m$, i. e. $x_i \neq x_j$ for all $i \neq j$. We will want to lay polynomials through given values, but we will now determine not only the values at those points, but also the first derivatives. That it, we set y_i a y'_i for all i. We are looking for a polynomial f which will satisfy these conditions on the values and derivatives.

5.20. 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. ○

5.21. Find any polynomial P satisfying

$$P(0) = 6, \quad P(1) = 4, \quad P(2) = 4, \quad P'(2) = 1.$$

○

5.22. 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. ○

5.23. Construct the natural cubic interpolation spline for the function

$$f(x) = |\,x\,|, \quad x \in [-1, 1]\,,$$

selecting the points $x_0 = -1$, $x_1 = 0$, $x_2 = 1$. ○

5.24. Construct the natural cubic interpolation spline for the points

$$x_0 = -3, \quad x_1 = 0, \quad x_2 = 3$$

and the values $y_0 = -3$, $y_1 = 0$, $y_2 = 3$. ○

5.25. 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. ○

5.26. Construct the complete (i. e., the derivatives at the marginal points are given) cubic interpolation spline for the points

$$x_0 = -3, \quad x_1 = -2, \quad x_2 = -1$$

and the values

$$y_0 = 0, \quad y_1 = 1, \quad y_2 = 2, \quad y_0' = 1, \quad y_2' = 1.$$

○

5.27. Construct the natural cubic interpolation spline for the function

$$y = \tfrac{1}{1+x^2},$$

selecting the points

$$x_0 = 0, \quad x_1 = 1, \quad x_2 = 3.$$

○

More problems concerning polynomial interpolation can be found at 315.

Analogously as in the case of interpolating the values only, we obtain the following system of $2(m+1)$ equations for the coefficients of the polynomial $f(x) = a_n x^n + \dots + a_0$:

$$\begin{aligned} a_0 + x_0 a_1 + \dots + (x_0)^n a_n &= y_0 \\ &\vdots \\ a_0 + x_m a_1 + \dots + (x_m)^n a_n &= y_m \\ a_1 + 2x_0 a_2 + \dots + n(x_0)^{n-1} a_n &= y_0' \\ &\vdots \\ a_1 + 2x_m a_2 + \dots + n(x_m)^{n-1} a_n &= y_m'. \end{aligned}$$

Again, we could verify that the choice $n = 2m+1$ makes the determinant of this system non-zero, and thus there will be exactly one solution. However, similarly to Lagrange polynomial, our polynomial f can be constructed straightaway. We just create a set of polynomials with values 0 or 1 (at both the derivatives and the values) in order to express the desired values as their linear combination. Verifying the following definition and proposition is left to the reader:

Hermite's interpolation polynomial

Hermite's interpolation polynomial is defined by fundamental Hermite's polynomials:

$$\begin{aligned} h_i^1(x) &= \left[1 - \frac{\ell''(x_i)}{\ell'(x_i)}(x - x_i)\right] (\ell_i(x))^2 \\ h_i^2(x) &= (x - x_i)\,(\ell_i(x))^2\,, \end{aligned}$$

where $\ell(x) = \prod_{i=1}^n (x - x_i)$. These polynomials satisfy:

$$\begin{aligned} h_i^1(x_j) &= \delta_i^j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \\ (h_i^1)'(x_j) &= 0 \\ h_i^2(x_j) &= 0 \\ (h_i^2)'(x_j) &= \delta_i^j \end{aligned}$$

and so Hermite's interpolation polynomial is given by the expression

$$f(x) = \sum_{i=1}^{k} \left(y_i h_i^1(x_i) + y_i' h_i^2(x_i) \right).$$

5.8. Examples of Hermite's polynomials. The simplest case is the one of prescribing the value and the derivative at one point. This fully determines a polynomial of degree one

$$f(x) = f(x_0) + f'(x_0)(x - x_0),$$

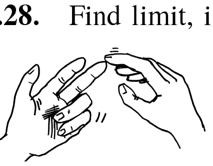
i. e. exactly the equation of the straight line given by the value and slope at the point x_0. When we set the values and the derivatives at two points, i. e. $y_0 = f(x_0)$, $y_0' = f'(x_0)$, $y_1 = f(x_1)$, $y_1' = f'(x_1)$ for two distinct points x_i, we still obtain an easily computable problem.

Let us look at it in a simple case when $x_0 = 0$, $x_1 = 1$. Then the matrix of the system and its inverse will be

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

B. Topology of the complex numbers and their subsets

5.28. Find limit, isolated, boundary, and interior points of the sets



$$\mathbb{N}, \quad \mathbb{Q}, \quad X = \{x \in \mathbb{R};\ 0 \le x < 1\} \text{ in } \mathbb{R}.$$

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

$$\delta_b := \inf_{n \in \mathbb{N}} |\, b - n\,| = \inf_{n \in \mathcal{O}_1(b) \cap \mathbb{N}} |\, b - n\,| > 0 \quad \text{for } b \in \mathbb{R} \setminus \mathbb{N}.$$

Therefore, $\mathcal{O}_{\delta_b}\,(b) \cap \mathbb{N} = \emptyset$, so no $b \in \mathbb{R} \setminus \mathbb{N}$ 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 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 $p/q + \sqrt{2}/10^n$ ($n \in \mathbb{N}$)

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.

$$\begin{aligned} f(x) = {} & (2y_0 - 2y_1 + y_0' + y_1')x^3 \\ & + (-3y_0 + 3y_1 - 2y_0' - y_1')x^2 + y_0'x + y_0. \end{aligned}$$

5.9. Spline interpolation. Similarly, 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 wanted polynomial leads to a unique interpolation. We will not delay ourselves with details here. Unfortunately, these interpolations do not solve the problems mentioned already in connection with the simple interpolation of values – complexity of the computations and instability. However, the usage of derivatives allows us to improve our methods:

As we have seen in the pictures demonstrating the instability of the interpolation by a single polynomial of sufficiently large degree, small local changes of the values dramatically affected the overall changes of the behavior of the resulting polynomial. Thus we may try to use small polynomial pieces of low degrees which we, however, must be able to link to one another properly.

The simplest case is to link each pair of adjacent points with a polynomial of degree at most one. This is also the most frequent way of displaying data. From the view of derivatives, this means that they will be constant on each of the segments and then will change in a leap.

A bit more sophisticated method is to prescribe the value and the derivative at each point, i. e. we will have four values for two points, which uniquely determines Hermite's polynomial of degree three, see above. This polynomial can then be used for all the values of the input variable between the marginal points $x_0 < x_1$. We talk about the *interval* $[x_0, x_1]$. Such a piecewise polynomial approximation has the property that the first derivatives will be compatible.

However, in practice, mere compatibility of the first derivatives is insufficient (for instance, with railway tracks), and furthermore, the values of the first derivatives are not always at our disposal. Thus we get the idea of making use of the values at the given points, and on the other hand to require equality of the first and second derivatives between the adjacent pieces of the cubic polynomials. This conditions yield the same number of equations and unknowns, and so the problem will be similarly solvable:

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 restriction of S on the interval $[x_{i-1}, x_i]$ is a polynomial S_i of degree at most three $i = 1, \ldots, n$
- $S_i(x_{i-1}) = y_{i-1}$ 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[1] for $n+1$ points consists of n cubic polynomials, i. e. we have $4n$ free parameters (the first condition from the

[1] The name comes from the meaning of a ruler used to draw smooth curves between points.

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. □

5.29. Determine the suprema and infima of the following sets in $\mathbb{R}$:

$$A = (-3, 0] \cup (1, \pi) \cup \{6\};\ \ B = \left\{ \frac{(-1)^n}{n^2};\ n \in \mathbb{N} \right\};\ \ C = (-9, 9) \cap \mathbb{Q}.$$

○

5.30. Find $\sup A$ and $\inf A$ for

$$A = \left\{ \frac{n + (-1)^n}{n};\ n \in \mathbb{N} \right\} \subset \mathbb{R}.$$

○

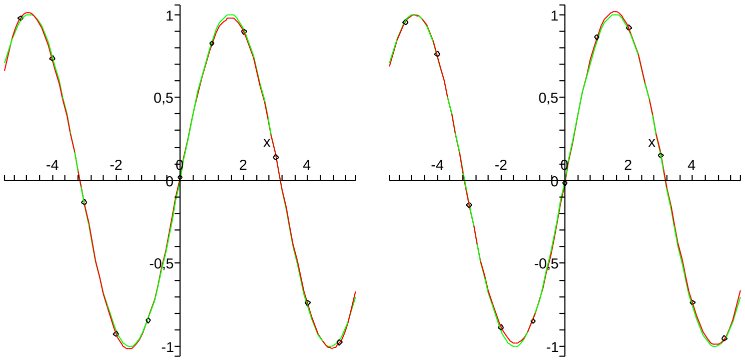
5.31. The following sets are given:

$$\mathbb{N} = \{1, 2, \dots, n, \dots\}, \quad \mathcal{M} = \left\{ -\frac{1}{n};\ n \in \mathbb{N} \right\},$$
$$\mathcal{J} = (0, 2] \cup [3, 5] \setminus \{4\}.$$

Determine $\inf \mathbb{N}$, $\sup \mathcal{M}$, $\inf \mathcal{J}$ and $\sup \mathcal{J}$ in $\mathbb{R}$. ○

definition). The other conditions then yield $2n + (n - 1) + (n - 1)$ more equalities, i. e. two parameters remain free. In practice, we often prescribe the values of the derivatives at the marginal points explicitly (the so-called *complete spline*), or assume they equal zero (this case is called a *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 mingle between adjacent intervals. However, with an appropriate order, one can obtain a matrix of the system such that all of its non-zero elements appear on three diagonals only. These matrices are nice enough to be solved in time proportional to the number of points, using a suitable numerical method. For comparison, let us look at interpolation of the same data as in the case of the Lagrange polynomial, now using splines:



2. Real number and limit processes

It is important to have a sufficiently large stock of functions with which we can express all usual dependencies. However, at the same time, the choice of the functions must be carefully restricted so that we would be able to build some universal and efficient tools for the work with them.

Actually, the first problem we have to solve is how to define the values of the functions at all. After all, all we can get with a finite number of multiplication and addition is polynomial functions and efficient manipulation can be done with rational numbers only. However, we cannot do with rational numbers even when looking for roots of quadratic polynomials as, for instance, $\sqrt{2}$ is not a rational number.

Thus our first step will be a thorough introducing of the so-called limit process, i. e. we will define precisely what it means that some values approach a certain value.

We can also notice that an important property of polynomials is their "continuous" dependency of their values on the input variable. Intuitively said, if we change x a little bit, the value of $f(x)$ also changes a bit only. On the other hand, this behavior is not possessed by piecewise constant functions $f : \mathbb{R} \to \mathbb{R}$ near the sudden "jumps". For instance, the so-called *Heaviside's function*

$$f(x) = \begin{cases} 0 & \text{for all } x < 0, \\ 1/2 & \text{for } x = 0, \\ 1 & \text{for all } x > 0 \end{cases}$$

has this type of "discontinuity" for $x = 0$.

Let us formalize these intuitive statements.

5.32. 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. ○

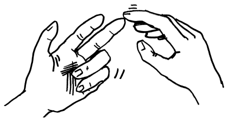
5.33. Find a subset X of the set $\mathbb{R}$ such that $\sup X \le \inf X$. ○

5.34. Find sets $A, B, C \subseteq \mathbb{R}$ such that

$$A \cap B = \emptyset,\ A \cap C = \emptyset,\ B \cap C = \emptyset,\ \sup A = \inf B = \inf C = \sup C.$$

○

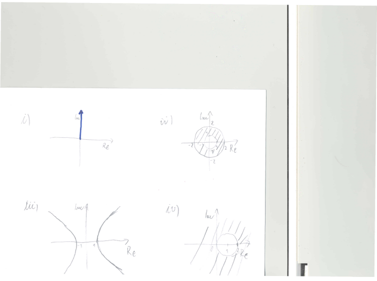
5.35. Mark the following sets in the complex plane:



i) $\{z \in \mathbb{C} |\ |z-1| = |z+1|\}$,
ii) $\{z \in \mathbb{C} |\ 1 \le |z-i| \le 2\}$,
iii) $\{z \in \mathbb{C} |\ \mathrm{Re}(z^2) = 1\}$,
iv) $\{z \in \mathbb{C} |\ \mathrm{Re}(\frac{1}{z}) < \frac{1}{2}\}$.

Solution.

- the imaginary axis,
- annulus around i,
- the hyperbola $a^2 - b^2 = 1$,
- exterior of the unit disc centered at 1.



□

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. 266.

5.10. Real numbers. So far, we have made do with algebraic properties of real numbers which claimed that $\mathbb{R}$ is a field. However, we have also used the relation of the standard (total) order of the real numbers, denoted "$\le$" (see the paragraph 1.38). 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 gradually 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 a so-called *ordered field.* Finally, the last axiom can be perceived as claiming that $\mathbb{R}$ is "sufficiently dense", i. e. there are no points missing between any points (like, for instance, $\sqrt{2}$ is missing in the rational numbers).

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 $1 \in \mathbb{R}$, $1 \ne 0$, such that for all $a \in \mathbb{R}$, $1 \cdot a = a$
(R8) for all $a \in \mathbb{R}$, $a \ne 0$, 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 $\le$ is a total order, i. e. reflexive, antisymmetric, transitive, and total on $\mathbb{R}$
(R11) for all $a, b, c \in \mathbb{R}$, $a \le b$ implies $a+c \le b+c$
(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 conception of a least upper bound (also called *supremum*) must be thoroughly introduced. It makes sense for any partially ordered set, i. e. a set with a (not necessarily total) ordering relation. We will also meet it later in algebraic contexts. Let us remind that at the general level, an ordering relation is any binary relation on a set which is reflexive, antisymmetric, and transitive; see the paragraph 1.38.

Supremum and infimum

Definition. Let us consider a subset $A \subset B$ in a partially ordered set B. An *upper bound* of the set A is any element $b \in B$ such that $b \ge a$ holds for all $a \in A$. Dually, we define the concept of a *lower bound* of the set A as an element $b \in B$ such that $b \le a$ for all $a \in A$.

The least upper bound of the set A, if it exists, is called its *supremum* and denoted by $\sup A$. Dually, the greatest lower bound, if it exists, is called an *infimum*; we write $\inf A$.

The last axiom of our table of properties of the real numbers thus claims that for every non-empty set A of real numbers, it is true that if there is a number a which is greater than or equal to all

5.36. 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) $\lim_{n\to-\infty} \frac{2^n-2^{-n}}{2^n+2^{-n}}$,

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 (5.21): $\forall n \in \mathbb{N} : \frac{\sqrt{4n^2}}{n} < \frac{\sqrt{4n^2+n}}{n} < \frac{\sqrt{4n^2+n+\frac{1}{16}}}{n}$. 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)

$$\begin{aligned}\lim_{n\to\infty} \sqrt{4n^2+n}-2n &= \lim_{n\to\infty} \frac{(\sqrt{4n^2+n}-2n)(\sqrt{4n^2+n}+2n)}{\sqrt{4n^2+n}+2n} \\ &= \lim_{n\to\infty} \frac{n}{\sqrt{4n^2+n}+2n} = \\ &= \lim_{n\to\infty} \frac{1}{\frac{\sqrt{4n^2+n}}{n}+2} = \frac{1}{4}.\end{aligned}$$

□

5.37. 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 $1+\varepsilon+\frac{\varepsilon^2}{4}$, especially $\sqrt[m]{c} < 1+\varepsilon+\frac{eps^2}{4}$. But then we have that

$$\sqrt[2m]{c} = \sqrt{\sqrt[m]{c}} < \sqrt{1+\varepsilon+\frac{\varepsilon^2}{4}} = 1+\frac{\varepsilon}{2} < 1+\varepsilon,$$

numbers $x \in A$, then there is a least number with this property. For instance, the choice $A = \{x \in \mathbb{Q}, x^2 < 2\}$ gives us the supremum $\sup A = \sqrt{2}$.

An immediate consequence of this axiom is also the existence of infima for any non-empty set of real numbers bounded from below. (It suffices to realize that changing the sign of all the numbers interchanges suprema and infima).

For the formal construction of our theory, we need to know whether the properties we demand from the real numbers are realizable, i. e. whether there is such a set $\mathbb{R}$ with the operations and ordering relation which satisfy the thirteen axioms. So far, we have constructed correctly only the rational numbers, which form an ordered field, i. e. satisfy the axioms (R1) – (R12), which can easily be verified.

Actually, the real numbers can not only be constructed, but the construction is, up to isomorphism, unique. However, for our need, we will do with an intuitive idea of the real line. We will focus on the existence and uniqueness later on.

5.11. The complex plane. Let us remind 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 a good 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 will work with them as we have already seen that extending some scalars to the complex numbers is highly advantageous for calculations, and sometimes even necessary.

There is an important operation on the complex numbers, the so-called *conjugation*. It is the reflection symmetry with respect to the line of real numbers, i. e. changing the sign of the imaginary part. We denote it by a bar over the number $z \in \mathbb{C}$:

$$\bar{z} = \operatorname{re} z - i \operatorname{im} z.$$

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 (zero). The square root of this non-negative real number is called the absolute value of the complex number z; we write

(5.3) $$|z|^2 = z \cdot \bar{z}.$$

The absolute value is also defined on any ordered field of scalars $\mathbb{K}$, we just define the *absolute value* $|a|$ as follows:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0. \end{cases}$$

Of course, it is true that for any numbers $a, b \in \mathbb{K}$,

(5.4) $$|a+b| \leq |a| + |b|.$$

This property is called the triangle inequality. It also holds for the absolute value of the complex numbers, which was defined above.

Especially for the field of rational numbers and the field of real numbers, which are subfields of the complex numbers, both definitions of the absolute value coincide.

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 $1/c$.

□

5.38. 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

$$n = (1+a_n)^n = 1 + \binom{n}{1}a_n + \binom{n}{2}a_n^2 + \cdots + a_n^n, \quad n \geq 2\,(n \in \mathbb{N}).$$

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 $c \geq 1$. □

5.39. 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

$$2^{\frac{1}{2}}\cdot 2^{\frac{1}{4}}\cdot 2^{\frac{1}{8}}\cdots 2^{\frac{1}{2^n}} = 2^{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}}.$$

Thus we get

$$\lim_{n\to\infty} \left(\sqrt{2}\cdot\sqrt[4]{2}\cdot\sqrt[8]{2}\cdots\sqrt[2^n]{2}\right) = \lim_{n\to\infty} 2^{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}}$$

$$= 2^{\lim\limits_{n\to\infty}\left(\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}\right)} = 2^{\sum\limits_{n=1}^{\infty}\frac{1}{2^n}}.$$

5.12. Convergence of a sequence. In the following paragraphs,

we will work with one of the number sets $\mathbb{K}$ of rational, real, or complex numbers. The absolute value thus must be understood in the corresponding context, and we should also bear in mind that the triangle inequality holds in all these cases.

We would like to formalize the notion of a sequence of numbers approaching a limit. Therefore, the key object of our interest will be sequences of numbers a_i, where the index i usually goes throughout the natural numbers. We will denote the sequences either loosely as $a_0, a_1, \dots$, or as infinite vectors $(a_0, a_1, \dots)$, or (similarly to the matrix notation) as $(a_i)_{i=1}^{\infty}$.

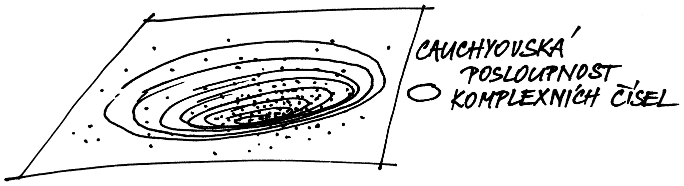
Cauchy sequences

Let us consider a sequence $(a_0, a_1, \dots)$ of elements of $\mathbb{K}$ such that for any fixed positive number $\varepsilon > 0$, it holds for all but finitely many terms a_i of the sequence that for all but finitely many terms a_j,

$$|a_i - a_j| < \varepsilon.$$

In other words, for any fixed $\varepsilon > 0$, there is an index N such that the above inequality holds for all $i, j > N$; i. e. the elements of the sequence are eventually arbitrarily close to each other. Such a sequence is called a *Cauchy sequence*.

Intuitively, we feel that either all but finitely many of the sequence's terms are equal (then $|a_i - a_j| = 0$ will hold from some index N on), or they "approach" some value. This is easily imaginable in the complex plane: choosing an arbitrarily small disc (with radius equal to ε), then, supposing we have a Cauchy sequence, it must be possible to put it into the complex plane in such a way that it covers all but finitely many of the elements of the infinite sequence a_i. We can imagine that the disc gradually shrinks to a single value a; see the picture.



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

We say that a sequence $(a_i)_{i=0}^{\infty}$ *converges* to a value a iff for any positive real number ε,

$$|a_i - a| < \varepsilon$$

holds for all but finitely many indeces i (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 $a_i \in \mathbb{K}$, $i = 0, 1, \dots$, converges to $a \in \mathbb{K}$, then for any fixed positive ε, we know that $|a_i - a| < \varepsilon$ for all i greater than a certain $N \in \mathbb{N}$. However, by the triangle inequality, we then get that for all pairs of indeces $i, j \geq N$, it is true that

$$|a_i - a_j| = |a_i - a_N + a_N - a_j| < |a_i - a_N| + |a_N - a_j| < 2\varepsilon.$$

Thus we have proved:

Lemma. *Every converging sequence is a Cauchy sequence.*

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

$$\lim_{n\to\infty}\left(\sqrt{2}\cdot\sqrt[4]{2}\cdot\sqrt[8]{2}\cdots\sqrt[2^n]{2}\right) = 2^1 = 2.$$

□

5.40. Determine

$$\lim_{n\to\infty}\left(\frac{1}{n^2}+\frac{2}{n^2}+\cdots+\frac{n-2}{n^2}+\frac{n-1}{n^2}\right).$$

○

5.41. Calculate

$$\lim_{n\to\infty}\frac{\sqrt{n^3-11n^2+2}+\sqrt[5]{n^7-2n^5-n^3}-n+\sin^2 n}{2-\sqrt[3]{5n^4+2n^3+5}}.$$

○

5.42. Determine the limit

$$\lim_{n\to\infty}\frac{n!+(n-2)!-(n-4)!}{n^{50}+n!-(n-1)!}.$$

○

5.43. 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

$$\lim_{n\to\infty}(x_n+y_n) = 1, \qquad \lim_{n\to\infty}\left(x_n\, y_n^2\right) = +\infty.$$

○

5.44. Determine the limit points of the sequence given by

$$a_n = \frac{(-1)^n\, 2n}{\sqrt{4n^2+5n+3}}, \qquad n \in \mathbb{N}.$$

○

5.45. Calculate

$$\limsup_{n\to\infty} a_n \quad \text{and} \quad \liminf_{n\to\infty} a_n$$

if

$$a_n = \frac{n^2+4n-5}{n^2+9}\,\sin^2\frac{n\pi}{4}, \qquad n \in \mathbb{N}.$$

○

5.46. Determine

$$\liminf_{n\to\infty}\left((-1)^n\left(1+\frac{1}{n}\right)^n+\sin\frac{n\pi}{4}\right).$$

○

However, in the field of rational numbers, it can easily happen that the corresponding value a does not exist even for a Cauchy sequence. For instance, the number $\sqrt{2}$ can be approached by rational numbers a_i with arbitrary accuracy, thereby obtaining a sequence converging to $\sqrt{2}$, but the limit is not rational.

Ordered fields of scalars in which every Cauchy sequence is converging are called *complete*. The following theorem proposes 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 value $a \in \mathbb{R}$.*

Proof. The terms of any Cauchy sequence form a bounded set since any choice of ε bounds all but finitely many of them. Let us define B as the set of those real numbers x for which $x < a_j$ holds for all but finitely many terms a_j of the sequence.

Apparently, B has an upper bound, and thus has a supremum as well, by (R13). Let us define $a = \sup B$. Now, having fixed some $\varepsilon > 0$, we choose N so that $|a_i - a_j| < \varepsilon$ for all $i, j \geq N$. Especially, $a_j > a_N - \varepsilon$ and $a_j < a_N + \varepsilon$ for all indeces $j > N$, and so $a_N - \varepsilon$ belongs to B, while $a_N + \varepsilon$ does not. Altogether, we get that $|a - a_N| \leq \varepsilon$, and thus

$$|a - a_j| \leq |a - a_N| + |a_N - a_j| \leq 2\varepsilon$$

for all $j > N$. However, this means that a is the limit of the considered sequence. □

Corollary. *Every Cauchy sequence of complex numbers z_i converges to a complex number z.*

Proof. Let us write $z_i = a_i + i\, b_i$. Since $|a_i - a_j|^2 \leq |z_i - z_j|^2$ and similarly for the values b_i, both sequences of real numbers a_i and b_i are Cauchy sequences. They converge to a and b, respectively, and we can easily verify that $z = a + i\, b$ is the limit of the sequence z_i. □

5.13. Remark. The previous discussion gives us a method for defining the real numbers. We proceed similarly to building the integers from the natural numbers (adding all additive inverses) and building the rational numbers from the integers (adding all multiplicative inverses of non-zero numbers). This time, we "complete" the rational numbers by all limits of Cauchy sequences.

It suggests itself to introduce a suitable equivalence relation on the set of all Cauchy sequences of rational numbers so that Cauchy sequences $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ are equivalent iff the distances $|a_i - b_i|$ converge to zero (this is the same as the condition that merging these sequences into a single sequence–for instance, the terms of the first sequence will become the odd terms of the resulting one and the terms of the second sequence will be the even ones–yields a Cauchy sequence as well). We will not verify that this relation is an equivalence in detail, neither will we define the operations and the ordering relation, nor will we prove that all of the axioms will indeed hold. Nevertheless, it is not difficult. Nor is proving the fact that the axioms (R1)–(R13) define the real numbers uniquely up to isomorphism (a bijective mapping preserving the algebraic operations as well as the ordering). We will return to this notes later.

5.47. Now let us proceed with limits of functions. The definition can be found at page 272.

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)

$$\lim_{x\to-\infty} \operatorname{arctg} \frac{1}{x}, \quad \lim_{x\to-\infty} \operatorname{arctg} x^4, \quad \lim_{x\to-\infty} \operatorname{arctg}\,(\sin x)\,.$$

Solution. Exercise (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

$$\lim_{x\to\pi/3} \sin x = \sin\frac{\pi}{3} = \frac{\sqrt{3}}{2}.$$

Exercise (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.

Exercise (c). By moving the limit inwards twice, the original limit transforms to

$$\left(\arccos\left(\lim_{x\to+\infty} \frac{1}{x+1}\right)\right)^3.$$

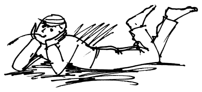
It can easily be shown that

$$\lim_{x\to+\infty} \frac{1}{x+1} = 0.$$

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

$$\lim_{x\to+\infty} \left(\arccos\frac{1}{x+1}\right)^3 = \left(\arccos\left(\lim_{x\to+\infty} \frac{1}{x+1}\right)\right)^3 = \left(\frac{\pi}{2}\right)^3.$$

5.14. Closed sets. Four our further work with the real or complex numbers, we will need to thoroughly understand the notions of closeness, boundedness, convergence, and so on. For any subset A of points in $\mathbb{K}$, we will be interested not only of the points belonging to $a \in A$, but also in the ones which can be approached by limits of sequences.



Limit points of a set

Let us consider a set A of points belonging to $\mathbb{K}$. A point $x \in \mathbb{K}$ is called a *limit point of the set A* iff there is a sequence $a_0, a_1, \dots$ of elements of A such that all its terms differ from x, yet its limit is x.

The limit points of a subset A of rational, real, or complex numbers are those numbers x which can be approached by such sequences of numbers lying in A which do not contain the point x itself. Let us notice that a limit point of a set may or may not belong to it.

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*. Let us notice that $d(x, A) = 0$ if and only if $x \in A$ or x is a limit point of A. (We suggest that the reader prove this in detail 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 such a set which coincides with its closure. Thus these are exactly those sets which contain all of its limit points as well. There is a typical example of a closed set: 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.

If either of the boundary values of the interval is missing, we write $a = -\infty$ (minus infinity) and similarly $b = +\infty$. Such closed intervals are denoted by $(-\infty, b]$, $[a, \infty)$, and $(-\infty, \infty)$.

The closed sets are exactly those which contain all they can "converge to". A closed set may 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.

We can easily verify that any intersection and any finite union of closed set is again a closed set. Indeed, if all of the points of some sequence belong to the considered intersection of closed sets, then they belong to each of the sets, and so do all the limit points. However, if we wanted to say the same about an arbitrary union, we would get in trouble: singleton sets are closed, but a sequence of points created from them may not be. On the other hand, if we restrict our attention to finite unions and consider a limit point of some 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 whole union.

Exercise (d). The function $y = \operatorname{arctg} x$ 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

$$\operatorname{arctg}\left(\lim_{x\to-\infty}\frac{1}{x}\right),\quad \operatorname{arctg}\left(\lim_{x\to-\infty}x^4\right),\quad \operatorname{arctg}\left(\lim_{x\to-\infty}\sin x\right).$$

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

$$\lim_{x\to-\infty}\operatorname{arctg}\frac{1}{x}=\operatorname{arctg}0=0,\quad \lim_{x\to-\infty}\operatorname{arctg}x^4=\lim_{y\to+\infty}\operatorname{arctg}y=\frac{\pi}{2}$$

and the last limit does not exist, either. □

5.48. Determine the limit

$$\lim_{x\to 0}\frac{1-\cos x}{x^2\sin(x^2)}.$$

Solution.

$$\begin{aligned}\lim_{x\to0}\frac{1-\cos x}{x^2\sin(x^2)} &= \lim_{x\to0}\frac{2\sin^2\left(\frac{x}{2}\right)}{x^2\sin(x^2)}=\\ &=\lim_{x\to0}\frac{\frac12\sin^2\left(\frac{x}{2}\right)}{\left(\frac{x}{2}\right)^2\sin(x^2)}=\\ &=\frac12\left(\lim_{x\to0}\frac{\sin\left(\frac{x}{2}\right)}{\frac{x}{2}}\right)^2\cdot\lim_{x\to0}\frac{1}{\sin^2(x^2)}=\frac12\cdot\infty=\infty.\end{aligned}$$

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 $\frac12\cdot\infty$ is meaningful (see the note after theorem 5.22), the original limit exists as well. If we split the original limit into the product

$$\lim_{x\to0}(1-\cos x)\cdot\lim_{x\to0}\frac{1}{x^2\sin(x^2)},$$

we would get the $0\cdot\infty$ type, which is an indeterminate form, but this tells us nothing about existence of the original limit. □

5.49. Determine the following limits:

i) $\lim_{x\to2}\frac{x-2}{\sqrt{x^2-4}}$, ii) $\lim_{x\to0}\frac{\sin(\sin x)}{x}$,
iii) $\lim_{x\to0}\frac{\sin^2 x}{x}$, iv) $\lim_{x\to0}e^{\frac{1}{x}}$.

Solution.

$$\text{i)}\ \lim_{x\to2}\frac{x-2}{\sqrt{x^2-4}}=\lim_{x\to2}\frac{x-2}{\sqrt{(x-2)(x+2)}}=\lim_{x\to2}\frac{\sqrt{x-2}}{\sqrt{x+2}}=\frac{0}{4}=0.$$

$$\text{ii)}\ \lim_{x\to2}\frac{x-2}{\sqrt{x^2-4}}\overset{(5.27)}{=}\lim_{y\to0}\frac{\sin y}{y}=1,$$

5.15. Open sets. There is another useful type of subsets 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 neighborhoods of points

An *open set* in $\mathbb{K}$ is a set whose complement is a closed set.

A *neighborhood of a point* $a\in\mathbb{K}$ is any open set $\mathcal{O}$ which contains a. If the neighborhood 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-neighborhood* of the point a.

Let us notice that for any set A, $a\in\mathbb{K}$ is a limit point of A if and only if every neighborhood of a contains at least one more point $b\in A$, $b\neq a$.

Lemma. *A set $A\subset\mathbb{K}$ of numbers is open if and only if with every point $a\in A$, an entire neighborhood of a belongs to A.*

Proof. Let A be an open set and $a\in A$. If there were no neighborhood of the point a inside A, there would be a sequence $a_n\notin A$, $|a-a_n|\le 1/n$. 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.

Now let us suppose that every $a\in A$ has an entire neighborhood of its lying in A. This naturally 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. □

From this lemma, it immediately follows that any union of open sets results in an open set, and further than any finite intersection of open sets is also an open set.

In the case of the real numbers, the δ–neighborhood 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.

5.16. Bounded and compact number sets. The closed and open sets are the basic concepts of *topology*. Without going into deeper connections, we have just made ourselves familiar with the *topology of the real line* and the *topology of the complex plane*. The following concepts will be extremely useful:

Bounded and compact sets

A set A of rational, real, or complex numbers is called *bounded* iff 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*.

Closed bounded intervals of real numbers are a typical example of compact sets.

Let us add further topological concepts that will allow us to express efficiently:

An *interior point* of a set A of real or complex numbers is such a point that one of its neighborhoods is contained in A.

A *boundary point* of a set A is such a point that all its neighborhoods are disjoint with neither A, nor its complement $\mathbb{K}\setminus A$. A boundary point of the set A may or may not belong to it.

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:

$$\lim_{x\to 0_+} e^{\frac{1}{x}} = e^{\lim_{x\to 0_+}\frac{1}{x}} = e^{\infty} = \infty,$$

$$\lim_{x\to 0_-} e^{\frac{1}{x}} = e^{\lim_{x\to 0_-}\frac{1}{x}} = e^{-\infty} = 0.$$

□

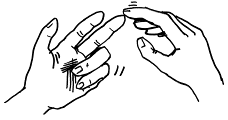
5.50. Calculate

(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 this 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 limits, supposing the particular limits exist and do not lead to one of the following expressions $\infty-\infty$, $0\cdot\infty$, $0/0$, ∞/∞, 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

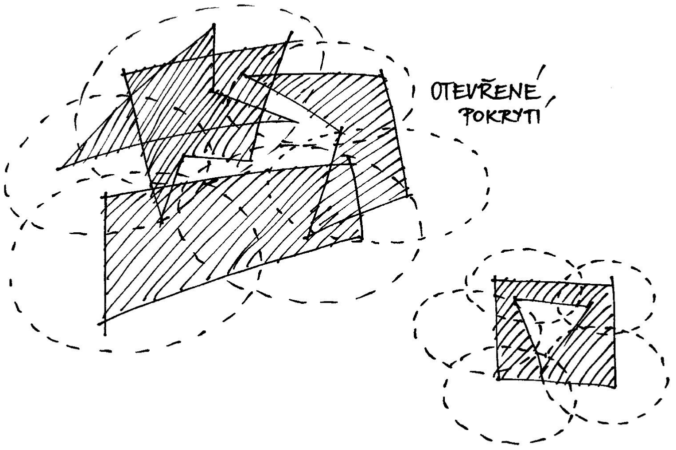
$$-\infty-\infty,\quad \frac{0}{3+\infty},\quad \frac{0}{(-\infty)^3-\infty}$$

can be called "determinate" (one can immediately determine the limit – they correspond to the values $-\infty$, 0, 0, respectively).

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

An *open cover* of a set A is such a system 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 neighborhood N of a satisfying $N\cap A = \{a\}$.



5.17. Theorem. *All subsets A of the real numbers satisfy:*

(1) a non-empty set A is open iff it is a union of countably (or finitely) many open intervals,

(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 iff every infinite sequence contained in it has a subsequence converging to a point in A,

(5) A is compact iff each of its open covers contains a finite subcover.

PROOF. (1) Apparently every open set is some union of neighborhoods of its points, i. e. of open intervals. So the question that remains is whether it suffices to take countably many of them. Thus we may try to select intervals which will be as "great" as possible. We will consider points $a,b\in A$ to be related iff the whole open interval $(\min\{a,b\},\max\{a,b\})$ is contained in A. Clearly, this relation is an equivalence (the open interval (a,a) is the empty set, which is contained in any set; symmetry and transitivity are apparent). The classes of this equivalence relation are intervals which are pairwise disjoint. Each of these intervals surely contains a rational number, and the obtained rational numbers are also pairwise distinct. However, there are only countably many rational numbers, so the statement is proved.

(2) It follows immediately from the definitions that no point can be both interior and boundary. 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 neighborhoods. Thus a is boundary.

(3) Suppose that $a\in A$ is boundary 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 that A is a compact set, i. e. both closed and bounded. Let us consider an infinite sequence of points $a_i\in A$. This set surely has both a supremum b and an infimum a (we could have taken any upper and lower bounds of the set A as well). Now let us cut 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. We will select this half and one of the terms contained in it;

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

$$f(x)^{g(x)} = \mathrm{e}^{\ln\left(f(x)^{g(x)}\right)} = \mathrm{e}^{g(x)\cdot \ln f(x)}.$$

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

$$\mathrm{e}^{\lim\limits_{x\to x_0}(g(x)\cdot \ln f(x))}.$$

Let us remind that either of these limits exists if and only if the other one exists. Further,

$$\begin{aligned}
\lim_{x\to x_0}(g(x)\cdot \ln f(x)) &= a \in \mathbb{R} &\Longrightarrow&\quad \lim_{x\to x_0} f(x)^{g(x)} = \mathrm{e}^{a},\\
\lim_{x\to x_0}(g(x)\cdot \ln f(x)) &= +\infty &\Longrightarrow&\quad \lim_{x\to x_0} f(x)^{g(x)} = +\infty,\\
\lim_{x\to x_0}(g(x)\cdot \ln f(x)) &= -\infty &\Longrightarrow&\quad \lim_{x\to x_0} f(x)^{g(x)} = 0.
\end{aligned}$$

Thus we can write

$$\lim_{x\to x_0} f(x)^{g(x)} = \mathrm{e}^{\lim\limits_{x\to x_0} g(x)\cdot \lim\limits_{x\to x_0} \ln f(x)},$$

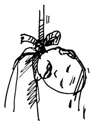
if both limits on the right-hand side exist and do not lead to the indeterminate form $0\cdot\infty$. It is not difficult to realize that this indeterminate form can only be obtained in three cases, corresponding to the remaining indeterminate forms 0^0, ∞^0, 1^∞, when we have, respectively, that

$$\begin{aligned}
\lim_{x\to x_0} f(x) &= 0 &\quad\text{and}\quad&& \lim_{x\to x_0} g(x) &= 0,\\
\lim_{x\to x_0} f(x) &= +\infty &\quad\text{and}\quad&& \lim_{x\to x_0} g(x) &= 0,\\
\lim_{x\to x_0} f(x) &= 1 &\quad\text{and}\quad&& \lim_{x\to x_0} g(x) &= \pm\infty.
\end{aligned}$$

then we cut the selected interval into halves. Again, we select such a half which contains infinitely many of the sequence's terms and select one of those points. By this procedure, we obtain a Cauchy sequence (you can prove this by yourselves; all you need is careful manipulation with some bound, similarly as above). However, we know that Cauchy sequences have limit points or are constant up to finitely many exceptions. Thus there is a subsequence with the wanted limit. >From the fact that A is closed, it follows that the obtained point 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 adjacent numbers would be at least 1, for instance. However, such a sequence of points in A cannot have a limit point at all.

(5) First, let us focus on the easier implication, i. e. let us suppose that every open cover contains a finite one and prove that A is both closed and bounded. Apparently, A can be covered by a countable union of intervals $I_n = (n-2, n+2)$, $n \in \mathbb{Z}$, and any choice of a finite subcover of them witnesses that A is bounded.



Now let us suppose that $a \in \mathbb{R} \setminus A$ is the limit point of a sequence $a_i \in A$, and further, let us assume that $|a - a_n| < \frac{1}{n}$ (otherwise we can select a subsequence satisfying this property). The sets

$$J_n = \mathbb{R} \setminus [a - \frac{1}{n}, a + \frac{1}{n}]$$

for all $n \in \mathbb{N}$, $n > 0$, are unions of two open intervals 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 $\mathbb{R} \setminus A$, including one of its neighborhoods, and thus it is not a limit point. Therefore, all of A's limit points must again lie in A. Hence A is closed as well.

The proof of the other implication is based upon the properties and existence of suprema. Let us suppose that A is compact and that an open cover $\mathcal{C}$ of A is given. >From the previous, it is apparent that A has a greatest element and a least element, which equal $b = \sup A$ and $a = \inf A$, respectively. Let us mark the "extreme" set for which we can choose a finite cover from $\mathcal{C}$, i. e. let us define the set

$$B = \{x \in [a,b],\ \text{there is a finite subcover } [a,x] \cap A\}.$$

Apparently, $a \in B$, so it is a non-empty set bounded from above. Therefore, it has a supremum c. Our task is to prove that, actually, $c = b$.

The reasoning may be a bit chaotic unless we draw a picture of the situation. However, the essence is simple: We know that $a \leq c \leq b$. Let us thus suppose, for a while, that $c < b$. Since $\mathbb{R} \setminus A$ is open, for $c \notin A$, there is a neighborhood of the point c contained in $[a,b]$ and, at the same time, disjoint with A. However, this would eliminate the possibility of $c = \sup B$.

So there remains the case $c \in A$, and there is a neighborhood $\mathcal{O}$ of the point c in the open cover $\mathcal{C}$. Let us choose points $p < c < q$ in $\mathcal{O}$. Again, there will be a finite cover for $[a,q] \cap A$. But this means that $q > c$ lies in B, which is impossible. The original choice of $c < b$ led to a contradiction, which proves the desired equality $b = c$. Now, with the help of a neighborhood of b lying in $\mathcal{C}$, we can find in $\mathcal{C}$ a finite cover for the whole of A. □

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)

$$\lim_{x\to x_0} f(x)^{g(x)} = \left(\lim_{x\to x_0} f(x)\right)^{\lim\limits_{x\to x_0} g(x)}.$$

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

$$\lim_{x\to+\infty} x^{-x} = \lim_{x\to+\infty}\left(x^x\right)^{-1} = 0.$$

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).

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. □

5.51. Calculate

$$\lim_{x\to+\infty}\frac{\sin x+\pi x^2}{2\cos x-1-x^2};$$

$$\lim_{x\to+\infty}\frac{3^{x+1}+x^5-4x}{3^x+2^x+x^2};$$

$$\lim_{x\to+\infty}\frac{4^x-8x^6-2^x-167}{3^x-45x-\sqrt{11}\pi^{x+12}};$$

$$\lim_{x\to+\infty}\frac{\sqrt{x}-\sin^3 x+x\operatorname{arctg} x}{\sqrt{1+2x+x^2}}.$$

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

$$\lim_{x\to+\infty}\frac{\frac{\sin x}{x^2}+\pi}{\frac{2\cos x-1}{x^2}-1} = \frac{0+\pi}{0-1} = -\pi.$$

5.18. 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 neighborhood of infinity is any interval (a, ∞). Similarly, any interval $(-\infty, a)$ is a neighborhood of $-\infty$. Further, we will extend the concept of a limit point so that ∞ is a limit point of a set $A \subset \mathbb{R}$ iff every neighborhood of ∞ has a non-empty intersection with it, i. e. if the set A is unbounded from above. Similarly for $-\infty$. We talk about infinite limit points, sometimes also called *improper limit points* of the set A.

"CALCULATIONS WITH INFINITIES"

We also introduce rules for calculation with the formally added values $\pm\infty$ and arbitrary "finite" numbers $a \in \mathbb{R}$:

$$\begin{aligned}
a+\infty &= \infty\\
a-\infty &= -\infty\\
a\cdot\infty &= \infty, \text{ if } a>0\\
a\cdot\infty &= -\infty, \text{ if } a<0\\
a\cdot(-\infty) &= -\infty, \text{ if } a>0\\
a\cdot(-\infty) &= \infty, \text{ if } a<0\\
\frac{a}{\pm\infty} &= 0, \text{ for all } a\neq 0.
\end{aligned}$$

The following definition covers many cases of limit processes and needs to be thoroughly understood. We will go through the particular cases in detail presently.

REAL AND COMPLEX LIMITS

Definition. Let us 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, let us consider a limit point x_0 of the set A (i. e. a real number or $\pm\infty$).

We sat that f has *limit* $a \in \mathbb{R}$ (or a complex limit $a \in \mathbb{C}$) at the point x_0 and write

$$\lim_{x\to x_0} f(x) = a$$

iff for every neighborhood $\mathcal{O}(a)$ of the point a, there is a neighborhood $\mathcal{O}(x_0)$ of the point x_0 such that for all $x \in A\cap(\mathcal{O}(x_0)\setminus\{x_0\})$, it holds that $f(x) \in \mathcal{O}(a)$.

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 has no occurrence in the definition, and the function f may even not be defined at this limit point (and in the case of an improper limit point, it cannot, of course)! We often talk about a *deleted neighborhood* $\mathcal{O}(x) \setminus \{x\}$ of those points where we are interested in the function values.

For now, we will not define improper limits of complex functions.

5.19. The most often cases of domains. Our definition of a limit covers several very dissimilar situations:

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)

$$\lim_{x\to+\infty} \frac{c}{x^\alpha} = 0, \ \lim_{x\to+\infty} \frac{x^\alpha}{x^\beta} = 0, \ \lim_{x\to+\infty} \frac{x^\beta}{a^x} = 0, \ \lim_{x\to+\infty} \frac{a^x}{b^x} = 0$$

for

$$c \in \mathbb{R}, \quad 0 < \alpha < \beta, \quad 1 < a < b.$$

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

$$\lim_{x\to+\infty} \frac{\sqrt{x} - \sin^3 x + x \operatorname{arctg} x}{\sqrt{1 + 2x + x^2}} = \lim_{x\to+\infty} \frac{x \operatorname{arctg} x}{\sqrt{x^2}} = \lim_{x\to+\infty} \operatorname{arctg} x$$
$$= \frac{\pi}{2}.$$

□

5.52. 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)

$$\frac{1}{(k-1)\,k} = \frac{1}{k-1} - \frac{1}{k},$$

(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 often 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 neighborhood $\mathcal{O}(a)$ 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 5.12). We have only added the possibility of infinite limits. We also say that the *sequence a_n converges to a*.

We can easily see from our definition for complex numbers that a sequence of complex values has limit a if and only if the real parts of a_i converge to $\operatorname{re} a$ and the imaginary parts converge to $\operatorname{im} a$.

(2) **Limits of functions at an interior point of an interval.** 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 neighborhoods of the limit point 0.

(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 left and from the right, 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 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 out 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.

5.20. Further examples of limits. (1) The limit of a complex function $f : A \to \mathbb{C}$ 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 definition of distances are neighborhoods of the points in the complex plane.

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

$$\frac{1}{\sqrt{n^2+1}}+\cdots+\frac{1}{\sqrt{n^2+n}}\geq\frac{1}{\sqrt{n^2+n}}+\cdots+\frac{1}{\sqrt{n^2+n}}=\frac{n}{\sqrt{n^2+n}},$$
$$\frac{1}{\sqrt{n^2+1}}+\cdots+\frac{1}{\sqrt{n^2+n}}\leq\frac{1}{\sqrt{n^2+1}}+\cdots+\frac{1}{\sqrt{n^2+1}}=\frac{n}{\sqrt{n^2+1}}$$

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

$$\lim_{n\to\infty}\frac{n}{\sqrt{n^2+n}}=\lim_{n\to\infty}\frac{n}{\sqrt{n^2}}=1,\quad\lim_{n\to\infty}\frac{n}{\sqrt{n^2+1}}=\lim_{n\to\infty}\frac{n}{\sqrt{n^2}}=1,$$

we also have that

$$\lim_{n\to\infty}\left(\frac{1}{\sqrt{n^2+1}}+\frac{1}{\sqrt{n^2+2}}+\cdots+\frac{1}{\sqrt{n^2+n}}\right)=1.$$

□

5.53. 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

$$\sqrt{1+x}+\sqrt{1-x}$$

Indeed, the membership into a δ–neighborhood of a complex value z is guaranteed by the real $(1/\sqrt{2})\delta$–neighborhoods 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 $x\in\mathbb{R}$, it holds that

$$\lim_{x\to x_0}f(x)=f(x_0).$$

Really, if $f(x)=a_nx^n+\cdots+a_0$, then the identity $(x_0+\delta)^k=x_0^k+k\delta x_0^{k-1}+\cdots+\delta^k$, substituted for $k=0,\dots,n$, gives that choosing a sufficiently small δ makes the values arbitrarily close to $f(x_0)$.

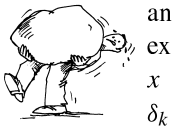
(3) Now, let us consider the following, quite awful, function defined on the whole real line

$$f(x)=\begin{cases}1 & \text{if } x\in\mathbb{Q}\\ 0 & \text{if } x\notin\mathbb{Q}.\end{cases}$$

It is apparent straight from the definition that this function has an (even one-sided) limit at no 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:[2]

$$f(x)=\begin{cases}\frac{1}{q} & \text{if } x=\frac{p}{q}\in\mathbb{Q},\ p,q \text{ relatively prime}\\ 0 & \text{if } x\notin\mathbb{Q}.\end{cases}$$



Choosing any point x, no matter whether rational or irrational, and a huge natural number m, then x will belong to exactly one of the intervals $(\frac{n}{m},\frac{n+1}{m})$ for some n (if $x=\frac{p}{q}$, we consider only coprime $m>q$). We set δ_k to be the minimum of the distances of the point x from the edges of these intervals for the considered m less than k. Of course, it always holds that $\delta_k<\frac{1}{k}$.

Now, let us consider some $\varepsilon>0$ and k such that $\frac{1}{k}<\varepsilon$. Then for all y in the deleted δ–neighborhood of the point x, we have either $f(y)=0$ (if it is an irrational value) or $f(y)<\frac{1}{r}$ for $r>k$ (if it is a rational value). In either case, we get that $|f(y)|<\varepsilon$.

Therefore, this function's limit is zero at all real points x. However, only at the irrational points, this limit equals the function value.

5.21. Theorem (The squeeze theorem)**.** *Let f, g, h be three real-valued functions with the same domain A and such that there is a deleted neighborhood of a limit point $x_0\in\mathbb{R}$ of the domain where*

$$f(x)\leq g(x)\leq h(x).$$

Then, supposing there are limits

$$\lim_{x\to x_0}f(x)=f_0,\quad\lim_{x\to x_0}h(x)=h_0$$

and $f_0=h_0$, the limit

$$\lim_{x\to x_0}g(x)=g_0$$

exists as well and it satisfies $g_0=f_0=h_0$.

[2] This function is called Thomae function after a German mathematician J. Thomae, 1840–1921.

and making use of the well-known formula $(a-b)(a+b)=a^2-b^2$. Thus we obtain

$$\begin{aligned}\lim_{x\to 0}\frac{\sqrt{1+x}-\sqrt{1-x}}{x}&=\lim_{x\to 0}\frac{(1+x)-(1-x)}{x\left(\sqrt{1+x}+\sqrt{1-x}\right)}\\&=\lim_{x\to 0}\frac{2}{\sqrt{1+x}+\sqrt{1-x}}=\frac{2}{\sqrt{1}+\sqrt{1}}=1.\end{aligned}$$

Similarly we can calculate

$$\begin{aligned}\lim_{x\to\pi/4}\frac{\cos x-\sin x}{\cos(2x)}&=\lim_{x\to\pi/4}\frac{(\cos x+\sin x)\,(\cos x-\sin x)}{(\cos x+\sin x)\cos(2x)}\\&=\lim_{x\to\pi/4}\frac{\cos^2 x-\sin^2 x}{(\cos x+\sin x)\cos(2x)}\\&=\lim_{x\to\pi/4}\frac{1}{\cos x+\sin x}=\frac{1}{\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}}=\frac{\sqrt{2}}{2}.\end{aligned}$$

The reduction was made thanks to the identity

$$\cos(2x)=\cos^2 x-\sin^2 x,\qquad x\in\mathbb{R}.$$

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 need 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

$$a=\sqrt[3]{x^2+2x+3},\qquad b=\sqrt[3]{x^2+2x+2}.$$

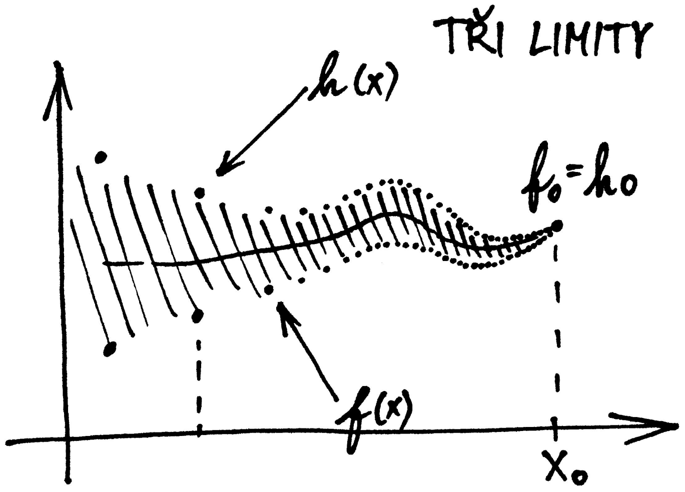
By this extension, we transform the original limit to for some polynomials P, Q. Let us emphasize that this really holds for all $n\in\mathbb{N}$. For $n=1$, one must realize that we set $\binom{1}{2}=0$ and that the polynomials P, Q may be constant zeros. So we get

$$\begin{aligned}(1+2nx)^n&=1+2n^2x+2n^3(n-1)\,x^2+P(x)\,x^3,\quad x\in\mathbb{R},\\(1+nx)^{2n}&=1+2n^2x+n^3(2n-1)\,x^2+Q(x)\,x^3,\quad x\in\mathbb{R}.\end{aligned}$$

Mere substitution and simple rearrangements give us

$$\begin{aligned}&\lim_{x\to 0}\frac{(1+2nx)^n-(1+nx)^{2n}}{x^2}=\\&\lim_{x\to 0}\frac{\left(2n^3(n-1)-n^3(2n-1)\right)x^2+(P(x)-Q(x))\,x^3}{x^2}=\\&\lim_{x\to 0}\left(-n^3+(P(x)-Q(x))\,x\right)=-n^3+0=-n^3.\end{aligned}$$

□



Proof. From the assumptions of the theorem, it follows that for any $\varepsilon>0$, there is a neighborhood $\mathcal{O}(x_0)$ of the point $x_0\in A\subset\mathbb{R}$ in which both $f(x)$ and $h(x)$ lie in the interval $(f_0-\varepsilon,\ f_0+\varepsilon)$, for all $x\neq x_0$. >From the condition $f(x)\le g(x)\le h(x)$, it follows that $g(x)\in(f_0-\varepsilon,\ f_0+\varepsilon)$ as well, so $\lim_{x\to x_0}g(x)=f_0$.

The presented reasoning can be gently modified for infinite limit values or for limits at infinite points x_0. It would be a good idea to think it through thoroughly! □

We can notice that this theorem allows us to calculate the limit for all types discussed above, i. e. limits of sequences, limits of functions at interior points, one-sided limits, and so on.

5.22. 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*

$$\lim_{x\to x_0}f(x)=a\in\mathbb{K},\qquad\lim_{x\to x_0}g(x)=b\in\mathbb{K}$$

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

$$\lim_{x\to x_0}(f(x)\cdot g(x))=a\cdot b,$$

(4) supposing $b\neq 0$, 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) Let us suppose that a and a' are two values of the limit $\lim_{x\to x_0}f(x)$. If $a\neq a'$, then there are disjoint neighborhoods $\mathcal{O}(a)$ and $\mathcal{O}(a')$. However, for sufficiently small neighborhoods of x_0, the values of f should lie in both the neighborhoods, which is a contradiction. Thus $a=a'$.

(2) Let us choose some neighborhood of $a+b$, for instance $\mathcal{O}_{2\varepsilon}(a+b)$. For a sufficiently small neighborhood of x_0 and $x\neq x_0$, both $f(x)$ and $g(x)$ will lie in ε–neighborhoods of the points a and b. Hence their sum will lie in the 2ε–neighborhood of the value $a+b$. The proposition is proved.

(3) Similarly to the above paragraph: we take $\mathcal{O}_{\varepsilon^2}(ab)$. For sufficiently small neighborhoods of x_0, the values of both f and

5.54. 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

$$\lim_{x\to x_0} f(x)^{g(x)} = \mathrm{e}^{\lim\limits_{x\to x_0}((f(x)-1)g(x))},$$

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

$$\begin{aligned}\lim_{x\to\pi/4}(\tan x-1)\tan(2x) &= \lim_{x\to\pi/4}\left(\frac{\sin x}{\cos x}-1\right)\frac{\sin(2x)}{\cos(2x)}\\ &= \lim_{x\to\pi/4}\frac{\sin x-\cos x}{\cos x}\cdot\frac{2\sin x\cos x}{\cos^2 x-\sin^2 x}\\ &= \lim_{x\to\pi/4}\frac{-2\sin x}{\cos x+\sin x} = \frac{-2\frac{\sqrt{2}}{2}}{\frac{\sqrt{2}}{2}+\frac{\sqrt{2}}{2}} = -1.\end{aligned}$$

Hence we have that

$$\lim_{x\to\pi/4}(\tan x)^{\tan(2x)} = \frac{1}{\mathrm{e}}.$$

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. □

5.55. 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

$$\sin x < x,\qquad x\in\left(0,\frac{\pi}{2}\right).$$

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)

$$x<\tan x,\qquad x\in\left(0,\frac{\pi}{2}\right).$$

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

$$\sin x < x < \frac{\sin x}{\cos x},\qquad x\in\left(0,\frac{\pi}{2}\right),$$

that is

$$1<\frac{x}{\sin x}<\frac{1}{\cos x},\qquad x\in\left(0,\frac{\pi}{2}\right),$$

$$1>\frac{\sin x}{x}>\cos x,\qquad x\in\left(0,\frac{\pi}{2}\right).$$

g will hit ε–neighborhoods of the values a and b. Therefore, their product will lie in the required ε^2–neighborhood.

(4) This is left as an exercise for the reader. □

Remark. If we look thoroughly at the presented proofs, we see that the statement of the theorem can be extended even to some infinite values of the limits of real-valued functions: Firstly, it must be the case that at least one of the limits is finite or that both limits share the same sign. Then it holds that the limit of the sum is the sum of the limits, with the conventions from 5.18. However, the case "$\infty-\infty$" is excluded.

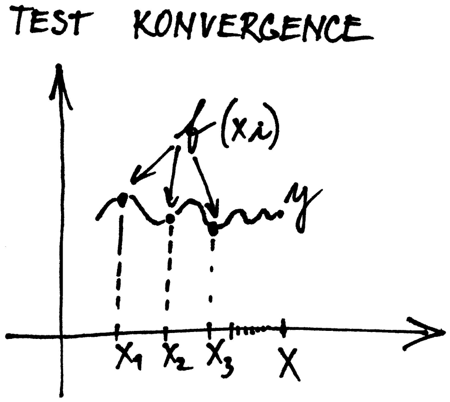
In the second case, one of the limits may be infinite, then the other one must be non-zero. Then, again, the limit of the product is the product of the limits. Now, the case "$0\cdot(\pm\infty)$" is excluded.

In the case of a quotient, we may have $a\in\mathbb{R}$ 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.

Let us emphasize that our 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.

For reasoning about limits, the following corollary of the definitions may be technically useful. It connects limits of sequences and of functions in general.

5.23. Corollary. *Let us consider a real or complex function f defined on a set $A\subset\mathbb{R}$ and a limit point x_0 of the set A. The function f has limit y at the point x_0 if and only if for every sequence of points $x_n\in A$ converging to, but different from x_0, the sequence of the values $f(x_n)$ has limit y.*



Proof. First, let us suppose that the limit of f at x_0 is y. Then for any neighborhood U of the point y, there must be a neighborhood V of the point x_0 such that for all $x\in V\cap A$, $x\neq x_0$, we have $f(x)\in U$. For every sequence $x_n\to x_0$ of points different from x_0, the terms x_n will lie in V for all n greater than a suitable N. Therefore, the sequences of values $f(x_n)$ will converge to y as well.

Now, let us suppose that the function f does not converge to y at $x\to x_0$. Then for some neighborhood U of the value y, there is a sequence of points $x_m\neq x_0$ in A which are closer to x_0 than $1/m$, and yet the value $f(x_m)$ does not belong to U. This way, we have constructed a sequence of points lying in A and different from x_0 for which the values $f(x_n)$ do not converge to y, thereby finishing the proof. □

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. □

5.56. Determine the limits

$$\lim_{n\to\infty} \left(\frac{n}{n+1}\right)^n, \quad \lim_{n\to\infty} \left(1 + \frac{1}{n^2}\right)^n, \quad \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^{n^2};$$

$$\lim_{x\to 0} \frac{\sin^2 x}{x}, \quad \lim_{x\to 0} \frac{x}{\sin^2 x}, \quad \lim_{x\to 0} \frac{\arcsin x}{x};$$

$$\lim_{x\to 0} \frac{3\tan^2 x}{5x^2}, \quad \lim_{x\to 0} \frac{\sin(3x)}{\sin(5x)}, \quad \lim_{x\to 0} \frac{\tan(3x)}{\sin(5x)};$$

$$\lim_{x\to 0} \frac{e^{5x} - e^{2x}}{x}, \quad \lim_{x\to 0} \frac{e^{5x} - e^{-x}}{\sin(2x)}.$$

Solution. When calculating these limits, we will use our knowledge of the following limits ($a \in \mathbb{R}$):

$$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a; \quad \lim_{x\to 0} \frac{\sin x}{x} = 1; \quad \lim_{x\to 0} \frac{e^x - 1}{x} = 1.$$

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

$$e^{-1} = \lim_{m\to\infty} \left(\frac{m}{m+1}\right)^m \cdot \lim_{m\to\infty} \frac{m}{m+1}.$$

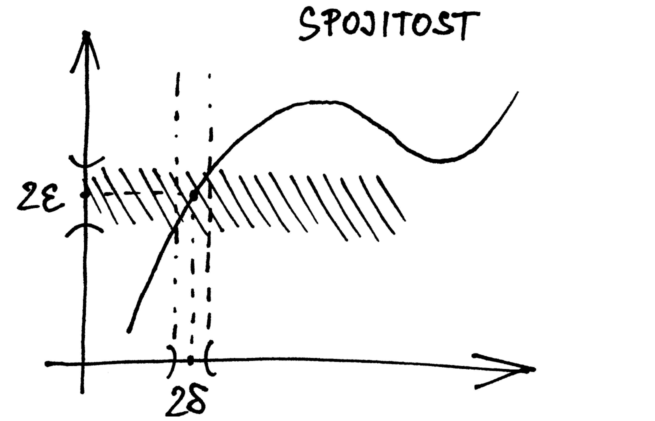
Now, we have prepared tools for a correct formulation of the property of continuity, with which we have dealt when talking about polynomials.

Continuity of functions

Definition. Let f be a real or complex function defined on an interval $A \subset \mathbb{R}$. We say that f is *continuous at a point* $x_0 \in A$ iff

$$\lim_{x\to x_0} f(x) = f(x_0).$$

The function f is said to be continuous on the set A iff it is continuous at every point $x_0 \in A$.



Let us 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. We say that the function is *right-continuous or left-continuous* at such a point. We have also seen that every polynomial is a continuous function on the whole $\mathbb{R}$, see 5.20(2). Further, we have met a function which is continuous at irrational real numbers only although it has limits at all rational points as well, see 5.20(4).

From the previous theorem 5.22 about limit properties, many of the following propositions immediately follow.

5.24. Theorem. *Let f and g be (real or complex) functions defined on an interval A 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 neighborhood of x_0 and is continuous at x_0.
(4) if a continuous function h is defined on an neighborhood of the value $f(x_0)$ of the real-valued function f, then the composite function $h \circ f$ is defined on an neighborhood of the point x_0 and is at x_0.

Proof. The statements (1) and (2) are apparent. We need to supplement the proof of (3). If $g(x_0) \neq 0$, then the entire ε–neighborhood of the number $g(x_0)$ does not contain zero for a sufficiently small $\varepsilon > 0$. >From the continuity of g, it follows that on a sufficiently small δ–neighborhood of the point x_0, g will be non-zero and the quotient f/g is thus well-defined there. However, then it will be continuous at x_0 by the previous theorem.

(4) Let us choose a neighborhood $\mathcal{O}$ of the value $h(f(x_0))$. >From the continuity of h, there is a neighborhood $\mathcal{O}'$ of the point

Clearly, the second limit is equal to 1. Changing the variables (replacing n with m), we can write the result

$$\mathrm{e}^{-1} = \lim_{n\to\infty} \left(\frac{n}{n+1}\right)^n .$$

Further, it holds that

$$\begin{aligned}\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^n &= \lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^{\frac{n^2}{n}} \\ &= \lim_{n\to\infty}\left(\left(1+\frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}} = \mathrm{e}^0 = 1\end{aligned}$$

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

$$\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)^{n^2} = \lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m = \mathrm{e}, \qquad \lim_{n\to\infty}\frac{1}{n} = 0$$

and the second one from

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = \mathrm{e}^{-1}, \qquad \lim_{n\to\infty} n = +\infty,$$

where we set $\mathrm{e}^{-\infty} = 0$ (this is a notation for $\lim_{x\to-\infty}\mathrm{e}^x = 0$, which is a determinate form).

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).

$f(x_0)$ which is mapped into $\mathcal{O}$ by h. The continuous function f maps some sufficiently small neighborhood of the point x_0 into the neighborhood $\mathcal{O}'$. However, this is the definition property of continuity, which finishes the proof. □

Now we can quite easily derive some basic connections between continuous mappings and the topology of the real numbers:

5.25. Theorem. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then*

(1) the inverse image $f^{-1}(U)$ of every open set U is an open set,
(2) the inverse image $f^{-1}(W)$ of every closed set W is a closed set,
(3) the image $f(K)$ of every compact set K is a compact set,
(4) f has both a maximum and a minimum on every compact set K.

Proof. (1) Let us consider a point $x_0 \in f^{-1}(U)$. There is a neighborhood $\mathcal{O}$ of the value $f(x_0)$ which is contained in U since U is open. However, then there is a neighborhood $\mathcal{O}'$ of the point x_0 which is mapped into $\mathcal{O}$ and thus belongs to the inverse image. Therefore, every point of the inverse image is interior, which finishes the proof.

(2) Let us 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 it. >From the continuity of f, it apparently follows that $f(x_i)$ converges to $f(x_0)$, and since W is closed, it must be that $f(x_0) \in W$. Clearly, all limit points of the inverse image of the set W are contained in W.

(3) Let us 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. We can select a finite cover from it, so it suffices to take finitely many of the corresponding images to cover the original set $f(K)$.

(4) Since the image of a compact set is again a compact set, the image must be bounded and it must contain both the supremum and the infimum. Hence it follows that these must also be the maximum and the minimum, respectively. □

5.26. 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 $[a, b]$.[3]

Proof. (1) First, let us 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 $y \notin f(A)$. This means that for open sets $B_1 = (-\infty, y)$ and $B_2 = (y, \infty)$, their inverse images $A_1 = f^{-1}(B_1) \subset A$ and $A_2 = f^{-1}(B_2) \subset A$ cover A. Again, these sets are open, disjoint, and have a non-empty intersection with A. Thus there must be a point $x \in A$ which does not lie in A_1 but is a limit point of this set. At the same time, it must lie in A_2, which is impossible for two disjoint open sets.

Thus we have 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 all must be below. Hence it follows that the image is again an interval. Let us notice that the marginal points of this interval may or may not lie in the image.

[3] This theorem is (especially in Czech literature) called Bolzano's theorem. Bernard Bolzano worked in Prague at the beginning of the 19th century.

We can immediately see that

$$\lim_{x\to 0}\frac{3\tan^2 x}{5\,x^2} = \lim_{x\to 0}\left(\frac{3}{5}\cdot\frac{\sin x}{x}\cdot\frac{\sin x}{x}\cdot\frac{1}{\cos^2 x}\right)$$
$$= \frac{3}{5}\cdot\lim_{x\to 0}\frac{\sin x}{x}\cdot\lim_{x\to 0}\frac{\sin x}{x}\cdot\lim_{x\to 0}\frac{1}{\cos^2 x} =$$
$$= \frac{3}{5}\cdot 1\cdot 1\cdot 1 = \frac{3}{5}.$$

By appropriate extension and substitution, we get

$$\lim_{x\to 0}\frac{\sin(3x)}{\sin(5x)} = \lim_{x\to 0}\left(\frac{\sin(3x)}{3x}\cdot\frac{5x}{\sin(5x)}\cdot\frac{3}{5}\right)$$
$$= \lim_{x\to 0}\frac{\sin(3x)}{3x}\cdot\lim_{x\to 0}\frac{5x}{\sin(5x)}\cdot\frac{3}{5}$$
$$= \lim_{y\to 0}\frac{\sin y}{y}\cdot\lim_{z\to 0}\frac{z}{\sin z}\cdot\frac{3}{5} = 1\cdot 1\cdot\frac{3}{5} = \frac{3}{5}.$$

Thanks to the previous result, it can easily be calculated that

$$\lim_{x\to 0}\frac{\tan(3x)}{\sin(5x)} = \lim_{x\to 0}\left(\frac{\sin(3x)}{\sin(5x)}\cdot\frac{1}{\cos(3x)}\right)$$
$$= \lim_{x\to 0}\frac{\sin(3x)}{\sin(5x)}\cdot\lim_{x\to 0}\frac{1}{\cos(3x)} = \frac{3}{5}\cdot 1 = \frac{3}{5}.$$

Similarly, we can determine

$$\lim_{x\to 0}\frac{\mathrm{e}^{5x}-\mathrm{e}^{2x}}{x} = \lim_{x\to 0}\left(\mathrm{e}^{2x}\,\frac{\mathrm{e}^{(5-2)x}-1}{(5-2)x}\,(5-2)\right)$$
$$= \lim_{x\to 0}\mathrm{e}^{2x}\cdot\lim_{x\to 0}\frac{\mathrm{e}^{3x}-1}{3x}\cdot 3$$
$$= \mathrm{e}^{0}\cdot\lim_{y\to 0}\frac{\mathrm{e}^{y}-1}{y}\cdot 3 = 1\cdot 1\cdot 3 = 3$$

and also

$$\lim_{x\to 0}\frac{\mathrm{e}^{5x}-\mathrm{e}^{-x}}{\sin(2x)} = \lim_{x\to 0}\left(\frac{\mathrm{e}^{5x}-1}{\sin(2x)} - \frac{\mathrm{e}^{-x}-1}{\sin(2x)}\right) =$$
$$\lim_{x\to 0}\left(\frac{\mathrm{e}^{5x}-1}{5x}\cdot\frac{2x}{\sin(2x)}\cdot\frac{5}{2} - \frac{\mathrm{e}^{-x}-1}{-x}\cdot\frac{2x}{\sin(2x)}\cdot\left(-\frac{1}{2}\right)\right) =$$
$$\lim_{x\to 0}\frac{\mathrm{e}^{5x}-1}{5x}\cdot\lim_{x\to 0}\frac{2x}{\sin(2x)}\cdot\frac{5}{2}$$
$$- \lim_{x\to 0}\frac{\mathrm{e}^{-x}-1}{-x}\cdot\lim_{x\to 0}\frac{2x}{\sin(2x)}\cdot\left(-\frac{1}{2}\right) =$$
$$\lim_{u\to 0}\frac{\mathrm{e}^{u}-1}{u}\cdot\lim_{z\to 0}\frac{z}{\sin z}\cdot\frac{5}{2} - \lim_{v\to 0}\frac{\mathrm{e}^{v}-1}{v}\cdot\lim_{z\to 0}\frac{z}{\sin z}\cdot\left(-\frac{1}{2}\right) =$$
$$\frac{5}{2}+\frac{1}{2} = 3.$$

□

5.57. 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

$$\lim_{x\to 0}\frac{\sin x}{x} = 1.$$

If the domain interval A contains one of its limit 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 as the image of a closed bounded interval (i. e. a compact set) must be a closed interval again. □

We will finish our introductory discussion by some more theorems which are useful tools for calculating limits.

5.27. Theorem (About the limit of a composite function)**.** *Let f, $g : \mathbb{R}\to\mathbb{R}$ be 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 from some deleted neighborhood 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 5.24(4). From the continuity of g at the point b, it follows that for any neighborhood V of the value $g(b)$, we can find a sufficiently small neighborhood 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 for some sufficiently small deleted neighborhood of the point a, which verifies the first statement.

Even if we cannot use the continuity of g at the point b, the previous reasoning will hold as well if we ensure that sufficiently small neighborhoods of the point a are mapped into a deleted neighborhood of the point b by the function f. □

5.28. Who is in the ZOO. We have begun to build our menagerie of functions with polynomials and functions which can be created from them "by parts". At the same time, we have derived many properties for a huge class of continuous functions. However, we do not have many practically manageable examples at our disposal (except for the polynomials). As another example, we will concentrate on the quotients of polynomials.

Let f and g be two polynomials which can take complex values as well (i. e. we admit expressions $a_n x^n + \cdots + a_0$ with complex coefficients $a_i\in\mathbb{C}$, but we allow to substitute real values only for the variable x).

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 at all real points x except for the roots of the polynomial g. Such functions are called *rational functions*. From the theorem 5.24, 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, supposing the point is a common root of both f and g and the multiplicity in f is at least as great as in g (in this case, extending the function's domain by this point and defining it to take the value of the limit there makes the functions continuous at the point as well),
- an infinite limit, supposing the one-sided infinite limits are equal,

Then, we get

$$\begin{aligned}\lim_{x\to 0}\frac{1-\cos(2x)}{x\sin x} &= \lim_{x\to 0}\frac{1-\left(\cos^2 x-\sin^2 x\right)}{x\sin x}\\ &= \lim_{x\to 0}\frac{\left(1-\cos^2 x\right)+\sin^2 x}{x\sin x}\\ &= \lim_{x\to 0}\frac{2\sin^2 x}{x\sin x} = \lim_{x\to 0} 2\frac{\sin x}{x} = 2;\end{aligned}$$

and

$$\begin{aligned}\lim_{x\to 0}\frac{1-\cos x}{x^2} &= \lim_{x\to 0}\left(\frac{1-\cos x}{x^2}\cdot\frac{1+\cos x}{1+\cos x}\right) = \lim_{x\to 0}\frac{1-\cos^2 x}{x^2\left(1+\cos x\right)}\\ &= \lim_{x\to 0}\frac{\sin^2 x}{x^2\left(1+\cos x\right)} = \left(\lim_{x\to 0}\frac{\sin x}{x}\right)^2\cdot\lim_{x\to 0}\frac{1}{1+\cos x}\\ &= \frac{1}{2}.\end{aligned}$$

Let us remark that we could also use the identity

$$1-\cos(2x) = 2\sin^2 x, \qquad x\in\mathbb{R}.$$

□

D. Continuity of functions

5.58. 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:

$$\begin{aligned}\lim_{x\to 0^-}(x-1)^{-\operatorname{sgn} x} &= \lim_{x\to 0^-}(x-1) = -1,\\ \lim_{x\to 0^+}(x-1)^{-\operatorname{sgn} x} &= \lim_{x\to 0^+}\tfrac{1}{x-1} = -1,\end{aligned}$$

whence $\lim\limits_{x\to 0}(x-1)^{-\operatorname{sgn} x} = -1$. However, the function value at 0 equals 1, so the examined function is not continuous at the point 0. Further, we have that

$$\begin{aligned}\lim_{x\to 1^-}(x-1)^{-\operatorname{sgn} x} &= \lim_{x\to 1^-}\tfrac{1}{x-1} = -\infty,\\ \lim_{x\to 1^+}(x-1)^{-\operatorname{sgn} x} &= \lim_{x\to 1^+}\tfrac{1}{x-1} = \infty.\end{aligned}$$

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. □

5.59. Without invoking the squeeze theorem, prove that the function

$$R(x) = \begin{cases} x, & x\in\left\{\frac{1}{n};\ n\in\mathbb{N}\right\};\\ 0, & x\in\mathbb{R}\setminus\left\{\frac{1}{n};\ n\in\mathbb{N}\right\}\end{cases}$$

is continuous at the point 0.

Solution. The function R is continuous at the point 0 if and only if

$$\lim_{x\to 0} R(x) = R(0) = 0.$$

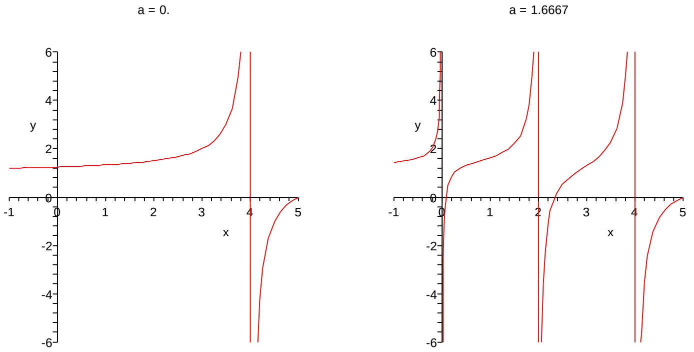
We will show that, by the definition of a limit, the examined limit equals 0. Using the "usual" notation, we have $a = 0$, $x_0 = 0$. Let $\delta > 0$ be arbitrary. For any $x\in(-\delta,\delta)$ we have that $R(x) = 0$,

- different one-sided infinite limits.

This situation is illustratively caught by the picture, which shows the values of the function

$$h(x) = \frac{(x-0.05a)(x-2-0.2a)(x-5)}{x(x-2)(x-4)}$$

for $a = 0$ (the left-hand picture thus displays the rational function $(x-5)/(x-4)$) and for $a = 5/3$.



5.29. Power and exponential functions. The polynomials are created by addition and multiplication of scalars and the simple power functions $x\mapsto x^n$ with natural exponents $n = 0, 1, 2, \dots$. The sense of the function $x\mapsto x^{-1}$, defined for all $x\neq 0$, is also obvious. Now, we will extend this definition to a general *power function* x^a with an arbitrary $a\in\mathbb{R}$.

We will use the properties of powers and roots, which we will consider to be a "matter of course". For a negative integer $-a$, we thus define

$$x^{-a} = (x^a)^{-1} = (x^{-1})^a.$$

Further, we would surely want the equality $b^n = x$ for $n\in\mathbb{N}$ to imply that b is the n–th root of x, i. e. $b = x^{\frac{1}{n}}$. It is necessary to verify that such b's always exist for positive real numbers x.

By factoring out $y_2 - y_1$ in $y_2^n - y_1^n$, we can easily see that the function $y\mapsto y^n$ is increasing for $y > 0$. Let us choose a number $x > 0$ 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 let us set $b = \sup B$. We already know that a power function with a natural exponent n is continuous, so we can easily verify that $b^n = x$. Indeed, surely $b^n\leq x$, and if the inequality were strict, we would find a number y such that $b^n < y^n < x$, which would imply that $b < y$, which contradicts the definition of a supremum.

Thus we have the power function correctly defined for all rational numbers $a = \frac{p}{q}$, $x^a = (x^p)^{\frac{1}{q}} = (x^{\frac{1}{q}})^p$.

Eventually, we can notice that for the values $a\in\mathbb{R}$ and $x > 1$, x^a is strictly increasing for rational a's. Therefore, we define

$$x^a = \sup\{x^y,\ y\in\mathbb{Q},\ y\leq a\}.$$

For $0 < x < 1$, we proceed analogously (one must be careful of the inequality signs) or we set $x^a = (\frac{1}{x})^{-a}$. For $x = 1$, we define $1^a = 1$ for any a.

Now, we have defined the power function $x\mapsto x^a$ for all $x\in[0,\infty)$ and $a\in\mathbb{R}$. However, we can consider 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 an *exponential function* with base c.

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

$$f(x) = \begin{cases} 1, & x \in \mathbb{Z}; \\ \frac{1}{q}, & x = \frac{p}{q} \in \mathbb{Q} \text{ for relatively prime } p, q \in \mathbb{Z} \text{ a } q > 1; \\ 0, & x \notin \mathbb{Q} \end{cases}$$

is also "often" called the Riemann function. □

5.60. By defining the values at the points -1 and 1, extend the function

$$f(x) = \left(x^2 - 1\right) \sin \frac{2x-1}{x^2-1}, \quad x \neq \pm 1 \; (x \in \mathbb{R})$$

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

$$\begin{aligned} f(-1) \;&:=\; \lim_{x \to -1} \left(\left(x^2 - 1\right) \sin \frac{2x-1}{x^2-1} \right), \\ f(1) \;&:=\; \lim_{x \to 1} \left(\left(x^2 - 1\right) \sin \frac{2x-1}{x^2-1} \right). \end{aligned}$$

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 $f(\pm 1) := 0$. □

5.61. 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

$$f(0) = -4, \quad \lim_{x \to +\infty} f(x) = \lim_{x \to +\infty} e^{2x} = +\infty.$$

The properties which we used when defining the power function and the exponential function $f(y) = c^y$, i. e. $c = f(1)$, can be summarized in a single inequality for any positive real x and real y:

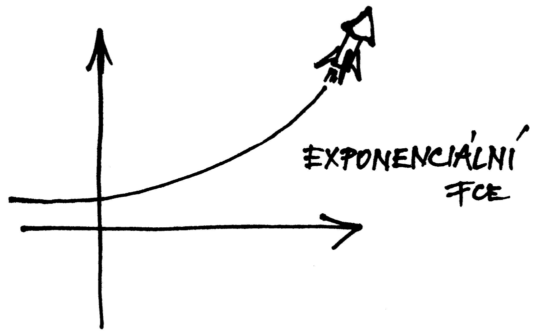
$$f(x+y) = f(x) \cdot f(y)$$

together with condition of continuity.

Indeed, for $y = 0$ we get that $f(0) = 1$, and hence $1 = f(0) = f(x-x) = f(x) \cdot (f(x))^{-1}$ and, eventually, for a natural number n, apparently $f(nx) = (f(x))^n$. Thus we have determined the values x^a for all $x > 0$ and $a \in \mathbb{Q}$. The continuity condition determines the function's values at the remaining points as well.

The exponential function especially satisfies the well-known formulas

$$a^x \cdot a^y = a^{x+y}, \quad (a^x)^y = a^{x \cdot y}. \tag{5.5}$$



5.30. Logarithmic functions. We have just seen that 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 a *logarithmic function with base* a. We write $\ln_a(x)$, and $\ln_a(a^x) = x$ is the defining property.

The equalities (5.5) 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 input values and are, on the whole domain, increasing for base $a > 1$ and decreasing for $0 < a < 1$. $\ln_a(1) = 0$ holds for every a.

We will see presently that there is an extremely important value of a, the so-called Euler's number e, see the paragraph 5.42. The function $\ln_e(x)$ is called the *natural logarithm* and denoted by $\ln(x)$ (i. e. omitting the base e).

3. Derivatives

When we were talking about polynomials, we already discussed how to describe the rate at which the function changes at a given point of its domain (see the paragraph 5.6). Back then, we examined the quotient (5.2), which expressed 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) increase Δx of the input variable. This reasoning is correct for any real or complex function f; we only have to properly work with the concept of a limit, instead of "intuitive decreasing" of Δx.

We introduce the definition of both proper and improper derivatives, i. e. we admit infinite values of the derivatives as well. We can notice that, unlike in the case of a mere limit of a function, now the function must be defined at the point x_0 at which we consider the derivative.

>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. □

5.62. At which points $x \in \mathbb{R}$ is the function

$$y = \cos\left(\operatorname{arctg}\left(\left|\,12x^{21} + 11\,\right| \cdot \frac{\mathrm{e}^{\cos(x+2)-x^3}}{-11 - x^{12}}\right)\right) + \sin\left(\sin\left(\sin x\right)\right)$$

(considering maximum domain) continuous? ○

5.63. 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 $-\pi, 0, 1, 2, 3, \pi$. ○

5.64. Extend the function

$$f(x) = \operatorname{arctg}\left(1 + \frac{5}{x^2}\right) \cdot \sin^2 x^5, \qquad x \in \mathbb{R} \setminus \{0\}$$

at $x = 0$ so that it is continuous at this point. ○

5.65. Find all $p \in \mathbb{R}$ for which the function

$$f(x) = \frac{\sin(6x)}{3x}, \quad x \in \mathbb{R} \setminus \{0\}; \qquad f(0) = p$$

is continuous at the origin. ○

5.66. 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 $\mathbb{R}$. ○

5.67. Calculate

$$\lim_{x \to 0+} \frac{\sin^8 x}{x^3}; \qquad \lim_{x \to -\infty} \frac{\sin^8 x}{x^3}.$$

○

5.68. 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 $a \neq 2$, 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

Derivative of a function of a real variable

5.31. 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, we say that the function f has *derivative a* at the point x_0. The value of the derivative is denoted by $f'(x_0)$ or $\frac{df}{dx}(x_0)$ or $a = \frac{d}{dx} f(x_0)$.

In accordance with the value of the defining limit, the derivative is also sometimes called *proper* or *improper*.

One-sided derivatives (i. e. left-sided derivatives and right-sided derivatives) are defined analogously in terms of the corresponding one-sided limits.

If a function has a derivative at a point x_0, we say the function is *differentiable* at x_0. A function which is differentiable at every point of a given interval is said to be differentiable on the interval.

Derivatives can be easily manipulated with, but we will have a lot of work correctly deriving the derivatives even of some already constructed functions. Therefore, a bit prematurely, we introduce a table of derivatives of several such functions. In the last column, you can find references to the corresponding paragraph where the result is proved. We can also notice that even though we are unable to express inverse functions to some of our functions by elementary means, we are nonetheless able to calculate their derivatives; see 5.35.

Derivatives of some functions

function	**domain**	**derivative**	
polynomials $f(x)$	whole $\mathbb{R}$	$f'(x)$ is again a polynomial	5.6
cubic splines $h(x)$	whole $\mathbb{R}$	only the first derivative of $h'(x)$ is continuous	5.9
rational functions $f(x)/g(x)$	whole $\mathbb{R}$ except for roots of g	rational functions: $\frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2}$	5.34
power functions $f(x) = x^a$	interval $(0, \infty)$	$f'(x) = ax^{a-1}$	??
exponential functions $f(x) = a^x$, $a > 0, a \neq 1$	whole $\mathbb{R}$	$f'(x) = \ln(a) \cdot a^x$	??
logarithmic functions $f(x) = \ln_a(x)$, $a > 0, a \neq 1$	interval $(0, \infty)$	$f'(x) = (\ln(a))^{-1} \cdot \frac{1}{x}$	??

From the formulation of the definition, we would anticipate that $f'(x_0)$ will allow us to approximate the function f by a straight line

$$y = f(x_0) + f'(x_0)(x - x_0).$$

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 $a \in [2, 6)$. □

5.69. 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. □

5.70. In $\mathbb{R}$, solve the equation

$$2^x + 3^x + 6^x = 1.$$

○

5.71. Determine whether the polynomial

$$x^{37} + 5x^{21} - 4x^9 + 5x^4 - 2x - 3$$

has a real root in the interval $(-1, 1)$. ○

E. Derivatives

First of all, let us show that the derivatives enlisted in the table of paragraph 5.31 are correct. We will derive them right from the definition of a derivative.

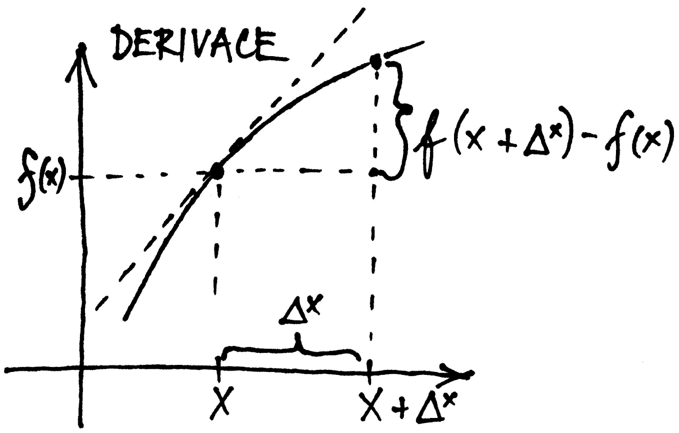
5.72. >From the definition, (see 5.31) find the derivatives of the functions x^n (x is the variable, n is a constant positive integer), $\sqrt{x}$, $\sin x$.

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.

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 the values $\psi(x)$ and the value $\psi(x_0)$ on a neighborhood of x_0 then says how much the slopes of the secant lines and the tangent line at the point 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 neighborhood $\mathcal{O}(x_0)$ and a function ψ which is continuous at x_0 and such that for all $x \in \mathcal{O}(x_0)$, it holds that*

$$f(x) = f(x_0) + \psi(x)(x - x_0).$$

Furthermore, then $\psi(x_0) = f'(x_0)$, and f itself is continuous at the point x_0.

Proof. First, let us suppose that $f'(x_0)$ is a finite derivative. If ψ is to exist, it is surely of the form

$$\psi(x) = (f(x) - f(x_0))/(x - x_0)$$

for all $x \in \mathcal{O} \setminus \{x_0\}$. On the other hand, we define the value at the point x_0 as $f'(x_0)$. Surely, then

$$\lim_{x \to x_0} \psi(x) = f'(x_0) = \psi(x_0)$$

as desired.

And 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 $\psi(x_0)$.

>From the expression of f in terms of continuous functions, it is apparent the f itself is continuous at the point x_0. □

5.32. Geometrical meaning of the derivative. The previous lemma can be illustrated geometrically, thereby getting another view at the derivative. It says that it can be determined whether the derivative exists from the graph of the function $y = f(x)$, i. e. the corresponding curve with coordinates x and y: the derivative exists if and only if the slope of the secant line going through the points $[x_0, f(x_0)]$ and $[x, f(x)]$ changes continuously. If so, the limit value of this slope is the value of the derivative.

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 neighborhood $\mathcal{O}(x_0)$ 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 $x \in \mathcal{O}(x_0)$, $x < x_0$.*

$$\begin{aligned}
(x^n)' &= \lim_{h\to 0}\frac{(x+h)^n - x^n}{h} = \lim_{h\to 0}\frac{\binom{n}{1}x^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \dots + h^n}{h}\\
&= nx^{n-1} + \lim_{h\to 0}\left(\binom{n}{2}x^{n-2}h + \binom{n}{3}x^{n-3}h^2 + \dots + h^{n-1}\right)\\
&= nx^{n-1},\\
(\sqrt{x})' &= \lim_{h\to 0}\frac{\sqrt{x+h}-\sqrt{x}}{h} = \lim_{h\to 0}\frac{(\sqrt{x+h}-\sqrt{x})(\sqrt{x+h}+\sqrt{x})}{h(\sqrt{x+h}+\sqrt{x})}\\
&= \lim_{h\to 0}\frac{h}{h(\sqrt{x+h}+\sqrt{x})} = \lim_{h\to 0}\frac{1}{\sqrt{x+h}+\sqrt{x}}\\
&= \frac{1}{2\sqrt{x}},\\
(\sin x)' &= \lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h}\\
&= \lim_{h\to 0}\frac{\sin x\cos h + \cos x \sin h - \sin x}{h}\\
&= \lim_{h\to 0}\frac{\cos x\sin h}{h} + \lim_{h\to 0}\frac{\sin x(\cos h - 1)}{h}\\
&= \cos x\cdot\lim_{h\to 0}\frac{\sin h}{h} - \lim_{h\to 0}\frac{2(\sin\frac{h}{2})^2}{h}\\
&= \cos x\cdot 1 + \lim_{t\to 0}\sin t\frac{\sin t}{t}\\
&= \cos x.
\end{aligned}$$

□

5.73. Differentiate:

i) $x\sin x$,
ii) $\frac{\sin x}{x}$,
iii) $\ln(x+\sqrt{x^2-a^2})$, $a\neq 0$, $|x|\geq|a|$,
iv) $\arctan\left(\frac{x}{\sqrt{1-x^2}}\right)$, $|x|\leq 1$,
v) x^x.

Solution. (i) By the formula for the derivative of a product (the Leibniz rule, see 5.33) 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 (5.34), 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 5.33). Setting $h(x) = \ln(x)$, $f(x) = x+\sqrt{x^2-a^2}$, we obtain

$$\begin{aligned}
\ln(x+\sqrt{x^2-a^2})' = h(f(x))' = h(f(x))\cdot f'(x) &= \frac{(x+\sqrt{x^2-a^2})'}{x+\sqrt{x^2-a^2}}\\
&= \frac{1+\frac{x}{x^2-a^2}}{x+\sqrt{x^2-a^2}},
\end{aligned}$$

On the other hand, if the derivative satisfies $f'(x_0) < 0$, then there is a neighborhood $\mathcal{O}(x_0)$ 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. Let us consider the former case. By the previous lemma, we have $f(x) = f(x_0) + \psi(x)(x-x_0)$ and $\psi(x_0) > 0$. However, since ψ is continuous at the point x_0, there must exist a neighborhood $\mathcal{O}(x_0)$ on which it holds that $\psi(x) > 0$. Then with increasing $x > x_0$, the value $f(x) > f(x_0)$ increases as well, and analogously for $x < x_0$.

The latter case (with a negative derivative) can be proved similarly. □

The functions that, for all points x of some neighborhood of a point x_0, satisfy $f(x) > f(x_0)$ if $x > x_0$ and $f(x) < f(x_0)$ if $x < x_0$ are called *increasing at the point* x_0. If the function is increasing at all points of a given interval, then it is said to be *increasing on the interval*. Of course, functions which are increasing on an interval satisfy $f(b) > f(a)$ for all $a < b$ from this interval.

Dually, a function is said to be *decreasing at a point* x_0 iff there is a neighborhood of the point x_0 such that $f(x) < f(x_0)$ if $x > x_0$ and $f(x) > f(x_0)$ if $x < x_0$ for all points x of the neighborhood. It is *decreasing on an interval* iff it is decreasing at every point of the interval.

Thus our corollary says that 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.

As an illustration of a simple usage of the connection between the derivatives and the properties of being an increasing (or decreasing) function, we can consider existence of inverses to polynomials. Since hardly any polynomials are exclusively increasing or decreasing functions, we cannot anticipate that there would be globally defined inverse functions to 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, when the equation

$$y = ax + b$$

gives that

$$x = \frac{1}{a}(y-b).$$

Similarly with a polynomial of degree two, the equation

$$y = ax^2 + bx + c$$

leads to the formula

$$x = \frac{-b\pm\sqrt{b^2-4a(c-y)}}{2a},$$

and thus the inverse (given by the above formula) exists only for those x which lie in either of the intervals $(-\infty, -\frac{b}{2a})$, $(-\frac{b}{2a}, \infty)$.

For the work with inverse functions to polynomials, we thus cannot do with the functions we have at our disposal now, so we obtain new additions to our menagerie.

where we used the chain rule once again when differentiating $\sqrt{x^2-a^2}$.

(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.

$$(x^x)' = \left((e^{\ln x})^x\right)' = (e^{x\ln x})'$$

$$= (x\ln x)'\cdot e^{x\ln x} = (1+\ln x)\cdot x^x$$

□

5.74. Find the derivative of the function $y = x^{\sin x}$, $x > 0$.

Solution. We have

$$\left(x^{\sin x}\right)' = \left(\mathrm{e}^{\sin x\ \ln x}\right)' = \mathrm{e}^{\sin x\ \ln x}\left(\cos x\ \ln x + \tfrac{\sin x}{x}\right) = x^{\sin x}\left(\cos x\ \ln x + \tfrac{\sin x}{x}\right).$$

□

5.75. For positive x, determine the derivative of the function

$$f(x) = x^{\ln x}.$$

○

Solution. $f'(x) = 2x^{\ln x-1}\cdot\ln x$.

5.76. For $x \in (0,\pi/2)$, calculate the derivative of the function

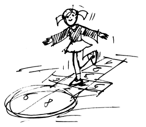
$$y = (\sin x)^{\cos x}.$$

○

Solution. $(\sin x)^{1+\cos x}\left(\cot^2 x - \ln(\sin x)\right)$.

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 5.32).

5.33. Formulae for calculation of derivatives. Now, we will introduce several basic facts about the calculation of derivatives. They will talk about how much the differentiation is 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.

In the case of a product, we have to be a bit more careful. 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, as we know, can be calculated as the sum of the products of the limits. Thus we can expect that the derivative of a product fg is given by the expression $fg'+f'g$, which is called *Leibniz rule*.

The derivative of a composite function is even more interesting: Let us consider a function

$$g = h\circ f,$$

where the domain of the function $z = h(y)$ contains the codomain of the function $y = f(x)$. Again, by writing out the increases, we obtain that

$$g' = \frac{\Delta z}{\Delta x} = \frac{\Delta z}{\Delta y}\frac{\Delta y}{\Delta x}.$$

Thus we may expect that the formula will be of the form

$$(h\circ f)'(x) = h'(f(x))f'(x).$$

Now we will provide correct formulations together with proofs:

Rules for differentiation

Theorem. *Let f and g be real or complex functions defined on a neighborhood 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 the point x_0 and

$$(cf)'(x_0) = c(f'(x_0)),$$

(2) the function $f+g$ has a derivative at the point x_0 and

$$(f+g)'(x_0) = f'(x_0) + g'(x_0),$$

(3) the function $f\cdot g$ has a derivative at the point x_0 and

$$(f\cdot g)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

5.77. Using differential, approximate arccotg 1, 02.

Solution. The differential of the function f having continuous derivative at the point x_0 is equal to

$$f'\left(x_0\right)\, dx = f'\left(x_0\right)\,\left(x - x_0\right).$$

The equation of the tangent to f's graph at the point $[x_0, f(x_0)]$ is then

$$y - f\left(x_0\right) = f'\left(x_0\right)\,\left(x - x_0\right).$$

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 we obtain the following formula for approximating the function value by its differential:

$$f(x) \approx f\left(x_0\right) + f'\left(x_0\right)\left(x - x_0\right).$$

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. □

5.78. Using differential, approximate arcsin 0, 497. ○

Solution. $\frac{\pi}{6} - \frac{2}{\sqrt{3}}\,0,003.$

5.79. Using differential, approximate

$$a := \operatorname{arctg} 1,02; \quad b := \sqrt[3]{70}.$$

○

Solution. $a \approx \frac{\pi}{4} + 0,01$; $b \approx 4,125$.

5.80. Using differential, approximate

(a) $\sin\left(\frac{29\pi}{180}\right)$;

(b) $\sin\left(\frac{46\pi}{180}\right)$.

○

Solution. (a) $\frac{1}{2} - \frac{\sqrt{3}\pi}{360}$; (b) $\frac{\sqrt{2}}{2} + \frac{\sqrt{2}\pi}{360}$.

5.81. 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

(4) Further, if h is a function defined on a neighborhood of the image $y_0 = f(x_0)$ and having a derivative at the point y_0, the composite function $h \circ f$ also has a derivative at the point 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) We will rewrite the quotient of the increases (which we have 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 expression for $x \to x_0$ gives the wanted result because f is continuous at the point x_0.

(4) By the lemma 5.31, there are functions ψ and φ which are continuous at the points x_0 and $y_0 = f(x_0)$ and they further satisfy

$$h(y) = h(y_0) + \varphi(y)(y - y_0), \;\; f(x) = f(x_0) + \psi(x)(x - x_0)$$

on some neighborhoods of x_0 and y_0. They also satisfy $\psi(x_0) = f'(x_0)$ and $\varphi(y_0) = h'(y_0)$. Then, it holds that

$$\begin{aligned} h(f(x)) - h(f(x_0)) &= \varphi(f(x))(f(x) - f(x_0)) \\ &= \varphi(f(x))\psi(x)(x - x_0) \end{aligned}$$

for x from the neighborhood of the point x_0. However, the product $\varphi(f(x))\psi(x)$ is a function which is continuous at x_0 and its value at the point x_0 is just the desired derivative of the composite function, again by the lemma 5.31. □

Derivative of a quotient

5.34. 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. First, we will proof the special case when $h(x) = x^{-1}$. From the definition of a derivative, we immediately get that

$$\begin{aligned} h'(x) &= \lim_{\Delta x \to 0} \frac{\frac{1}{x+\Delta x} - \frac{1}{x}}{\Delta x} = \lim_{\Delta x \to 0} \frac{x - x - \Delta x}{\Delta x(x^2 + x\Delta x)} \\ &= \lim_{\Delta x \to 0} \frac{-1}{x^2 + x\Delta x} \end{aligned}$$

and from the formulae for manipulations with limit, it follows that

$$h'(x_0) = -x^{-2}.$$

Now, the chain rule says that

$$(g^{-1})' = -g^{-2} \cdot g',$$

and, eventually, by the product rule, we obtain

$$(f/g)' = (f \cdot g^{-1})' = f'g^{-1} - fg^{-2}g' = \frac{f'g - gf'}{g^2}.$$

□

given point, thus we get the condition

$$\left.\frac{2-\ln(cx)}{2\sqrt{x}}\right|_{x=1} = 2, \text{ that is } \quad 2-\ln(c)=4,$$

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. □

5.82. Determine whether there is a point in the interval $(0, 4)$ such that the tangent line at this point to the polynomial $x\,(x-4)^5$ is parallel to the x-axis. ○

Solution. Yes, there is.

5.83. Let $p \in (0, +\infty)$. Write the equation of the tangent to the parabola $2py = x^2$ at a general point $[x_0, ?]$. ○

Solution. $y = \frac{x_0}{p}\,x - \frac{x_0^2}{2p}$.

5.84. 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. ○

Solution. $y = 2x$.

5.85. Find the equations of the tangent and normal lines to the curve

$$y = (x+1)\,\sqrt[3]{3-x}, \qquad x \in \mathbb{R}$$

at the point $[-1, 0]$. ○

Solution. $y = \sqrt[3]{4}\,(x+1)$; $y = -\frac{\sqrt[3]{2}}{2}\,(x+1)$.

5.86. Let the function

$$y = \frac{\ln(2x^3+4x^2-x)}{1+x}, \qquad x \in \left(\tfrac{1}{2}, +\infty\right)$$

be given. Determine the equations of the tangent and normal lines to the graph of this function at the point $[1, ?]$. ○

Solution. $y - \frac{\ln 5}{2} = \left(\frac{13}{10} - \frac{\ln 5}{4}\right)(x-1)$; $y - \frac{\ln 5}{2} = \frac{20}{5\ln 5 - 26}(x-1)$.

5.87. At which points is the tangent to the parabola

$$y = 2 + x - x^2, \qquad x \in \mathbb{R}$$

parallel to the x-axis? ○

Solution. $\left[\frac{1}{2}, 2\frac{1}{4}\right]$.

5.88. Determine the equations of the tangent line t and the normal line n to the graph of the function

$$y = \sqrt{x^2 - 3x + 11}, \qquad x \in \mathbb{R}$$

at the point $[2, ?]$. Further, determine all points at which the tangent is parallel to the x-axis. ○

Solution. $t: y = \frac{x}{6} + \frac{8}{3}$; $n: y = -6x + 15$; $\left[\frac{3}{2}, \sqrt{\left(\frac{3}{2}\right)^2 - 3\,\frac{3}{2} + 11}\right]$.

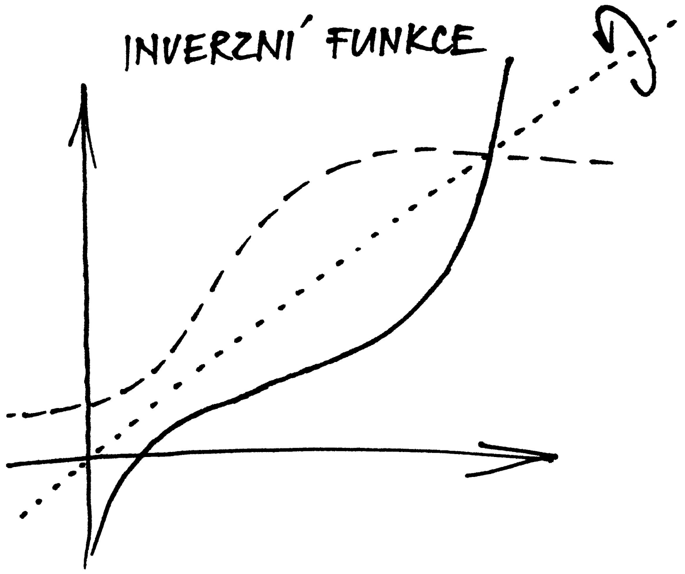
5.89. What is the angle between the x-axis and the graph of the function $y = \ln x$? (We mean the angle between the tangent line and the

5.35. Derivatives of inverse functions. In the paragraph 1.36,

while talking about relations and mappings in general, we have defined the concept of an *inverse function.* 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

$$f^{-1} \circ f = \mathrm{id}_{\mathbb{R}}, \quad f \circ f^{-1} = \mathrm{id}_{\mathbb{R}},$$

and the other one then holds as well. 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 we can see from the picture, the graph of the inverse function is obtained simply by interchanging the axes of the input and output variables.



If we knew that the inverse $y = f^{-1}(x)$ of a differentiable function $x = f(y)$ is also differentiable, then the chain rule would immediately yield

$$1 = (\mathrm{id})'(x) = (f \circ f^{-1})'(x) = f'(y) \cdot (f^{-1})'(x),$$

so we obtain the formula (apparently, $f'(y)$ must be non-zero in this case)

Derivative of an inverse function

$$(f^{-1})'(x) = \frac{1}{f'(y)}. \tag{5.6}$$

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 how we can calculate the derivatives of inverse functions:

Theorem. *If f is a real-valued function differentiable at a point x_0 and such that $f'(x_0) \neq 0$, then there is a function f^{-1} defined on some neighborhood of the point and such that (5.6) holds.*

positive x-axis in the "positive sense of rotation", ie. counterclockwise.) ○

Solution. $\pi/4$.

5.90. 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 $[1, 1]$. ○

Solution. $y = 2 - x$; $y = x$.

5.91. Prove the following:

$$\tfrac{x}{1+x} < \ln{(1+x)} < x \qquad \text{for all } x > 0.$$

○

Solution. The inequalities follow, for instance, from the mean value theorem (attributed to Lagrange) applied to the function $y = \ln{(1+t)}\,,\, t \in [0, x]$.

F. Extremal problems

The simple observation 5.32 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.

5.92. 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

$$f(x) = \sqrt{(x-1)^2 + (x^2-2)^2}, \quad x \in \mathbb{R}.$$

Apparently, 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

$$\tfrac{1-\sqrt{3}}{2} \quad \text{and} \quad \tfrac{1+\sqrt{3}}{2}.$$

As the function g is a polynomial (differentiable on the whole domain), from the geometrical sense of the problem, we get

Proof. First, let us realize that the request that the derivative at x_0 be non-zero means that our function f is either increasing or decreasing on some neighborhood of the point; see the corollary 5.32. Thus there exists an inverse function defined on some neighborhood. Since a continuous function maps a closed bounded interval onto a closed interval; the image $f(U)$ of any open set U contained in the domain of f is open as well. Then, by the definition of continuity, the inverse function is continuous, too.

To prove our proposition, it now suffices to carefully read through the proof of the fourth statement of the theorem 5.33. We only choose f for h and f^{-1} for f, and we know that the composite function is differentiable instead of supposing existence of the derivatives of both the functions (and we know that the composite is the identity function): Indeed, by the lemma 5.31, there is a function ψ continuous at the point y_0 such that

$$f(y) - f(y_0) = \varphi(y)(y - y_0),$$

on some neighborhood of y_0. Further, it satisfies $\varphi(y_0) = f'(y_0)$. However, then the substitution $y = f^{-1}(x)$ gives that

$$x - x_0 = \varphi(f^{-1}(x))(f^{-1}(x) - f^{-1}(x_0)),$$

for all x lying in some neighborhood $\mathcal{O}(x_0)$ of the point x_0. Further, we have $f^{-1}(x_0) = y_0$, and since f is either strictly increasing or strictly decreasing, we get that $\varphi(f^{-1}(x)) \neq 0$ for all $x \in \mathcal{O}(x_0) \setminus \{x_0\}$. Thus we can write

$$\frac{f^{-1}(x) - f^{-1}(x_0)}{x - x_0} = \frac{1}{\varphi(f^{-1}(x))} \neq 0,$$

for all $x \in \mathcal{O}(x_0) \setminus \{x_0\}$. The right-hand side of this expression is continuous at the point x_0 and the limit equals

$$\frac{1}{\varphi(f^{-1}(x_0))} = \frac{1}{f'(y_0)}.$$

Therefore, the limit of the left-hand side exists as well and equals the expression, i. e.

$$(f^{-1})'(x_0) = \frac{1}{f'(y_0)}$$

exists. □

5.36. Derivatives of power, exponential, and logarithmic functions. As an illustrating example of calculating the derivative of an inverse function, we will determine $(\ln_e)'$. We will use the formula $(e^x)' = e^x$ although we have not proved it yet. From the definition of the natural logarithm,

$$e^{\ln x} = x,$$

so we can easily calculate:

$$(\ln)'(y) = (\ln)'(e^x) = \frac{1}{(e^x)'} = \frac{1}{e^x} = \frac{1}{y}. \tag{5.7}$$

The formula

$$(x^a)' = ax^{a-1} \tag{5.8}$$

for differentiating a general power function can also be derived using the derivatives of exponential and logarithmic functions:

$$(x^a)' = (e^{a \ln x})' = e^{a \ln x}(a \ln x)' = ax^{a-1}.$$

$$x_A = \tfrac{1+\sqrt{3}}{2}.$$

□

5.93. Consider an isosceles triangle with base length b and height (above the base) h. Inscribe the rectangle having the greatest possible area into it (one of the rectangle's sides will lie on the triangle's base). Determine the area S of this rectangle.

Solution. To solve this problem, it suffices to consider the problem of inscribing the largest rectangle into a right triangle with legs of lengths $b/2$ and h so that two of the rectangle's sides lie on the legs of the triangle. Thus we reduce the problem to maximizing the function

$$f(x) = x\left(h - \tfrac{2hx}{b}\right)$$

on the interval $I = [0, b/2]$. Since we have that

$$f'(x) = h - \tfrac{4hx}{b} \qquad \text{for all } x \in I$$

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 $S = hb/4$. □

5.94. 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. □

5.95. 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 $[-\frac{c}{a}, 0]$, $[0, -\frac{c}{b}]$ and the area of the triangle whose vertices are these two points and the origin is $S = \frac{c^2}{2ab}$. 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

Now, let us focus on the derivative of the exponential function $f(x) = a^x$. If the derivative of a^x exists at all points x, it will surely hold that

$$f'(x) = \lim_{\Delta x \to 0} \frac{a^{x+\Delta x} - a^x}{\Delta x} = a^x \lim_{\Delta x \to 0} \frac{a^{\Delta x} - 1}{\Delta x} = f'(0)a^x.$$

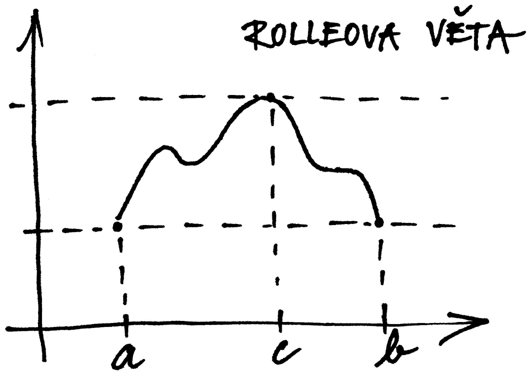
On the other hand, if the derivative at zero exists, then this formula guarantees the existence of the derivative at any point of the domain and also determines its value. At the same time, we verified the validity of the formula for the one-sided derivatives.

Unfortunately, it will take us some time to verify (see 5.43, ||i||, and 6.43) that the derivatives of exponential functions indeed exist. We will also see that there is an especially important base e, the so-called Euler's number, for which the derivative at zero equals one. What we can do now is to notice that the exponential functions are special in the way that their derivatives are proportional (with a constant coefficient) to their values:

$$(a^x)' = (\mathrm{e}^{\ln(a)x})' = \ln(a)(\mathrm{e}^{\ln(a)x}) = \ln(a) \cdot a^x.$$

5.37. Mean value theorems. Before we continue our journey of building new functions, we will derive several simple statements about the derivatives. The meaning of all of them is intuitively clear from the pictures and the proofs only follow the visual imagination.

Theorem. *Let a function $f : \mathbb{R} \to \mathbb{R}$ be 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 reaches both a maximum and a minimum there. If the maximum and the minimum shared the value $f(a) = f(b)$, it would mean that the function f is constant, and thus its derivative is zero at all points of the interval (a, b). Therefore, let us suppose that at least one of the maximum and the minimum is different and let it occur at an interior point c. Then it is impossible to have $f'(c) \neq 0$ because then the function f would be either increasing or decreasing at the point c (see 5.32) and so it would take both lower and higher values than $f(c)$ at a neighborhood of the point c. □

The above theorem is called *Rolle's theorem.*It immediately implies the following corollary, known as *Lagrange's mean value theorem.*

5.38. Theorem. *Let a function $f : \mathbb{R} \to \mathbb{R}$ be continuous on an interval $[a, b]$ and differentiable at all points inside this interval.*

thus $S = \frac{2}{3x_Ty_T}$. Further, in the first quadrant, we have that $x_T, y_T > 0$. To minimize this area means to maximize the product $x_Ty_T = x_T\sqrt{2-3x_T^2}$, which is (in the first quadrant) the same as to maximize $(x_Ty_T)^2 = x_T^2(2-3x_T^2) = -3(x_T^2-\frac{1}{3})^2+\frac{1}{3}$. Hence, the wanted minimum is at $x_T = \frac{1}{\sqrt{3}}$. The tangent's equation is $\sqrt{3}x + y = 2$ and the triangle's area is $S_{min} = \frac{2\sqrt{3}}{9}$. □

5.96. 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 1.34). So we minimize the determinant:

$$\begin{vmatrix} -2+t & t \\ -t^2-2t & -1+t^2 \end{vmatrix} = 2t^3 - t + 2.$$

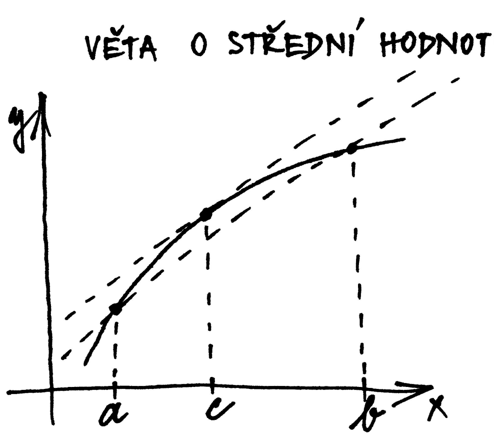
The derivative is $6t^2 - 1$, so the extrema occur at $t = \pm\frac{1}{\sqrt{6}}$. Thanks to considering non-negative time only, we are interested in $t = \frac{1}{\sqrt{6}}$. The second derivative of the considered function is positive at this point, thus the function has a local minimum there. Further, 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. □

5.97. 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

Then there is a number $c \in (a, b)$ such that

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$



Proof. The proof is a simple record 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 picture). The equation of our 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 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 $h'(c) = 0$. □

The mean value theorem can also be written in the form:

$$f(b) = f(a) + f'(c)(b-a). \tag{5.9}$$

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 existence of a tangent line parallel to the secant line going through the marginal points is described by the so-called *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. Again, we rely on Rolle's theorem. Thus we set

$$h(t) = (f(b)-f(a))g(t) - (g(b)-g(a))f(t).$$

Now $h(a) = f(b)g(a) - f(a)g(b)$, $h(b) = f(b)g(a) - f(a)g(b)$, so there is a number $c \in (a, b)$ such that $h'(c) = 0$. Since $g'(c) \neq 0$ we get just the desired formula. □

A reasoning similar to the one in the above proof leads to a supremely useful tool for calculating limits of quotients of functions. The theorem is known as *l'Hospital's rule*:

the following parametric equations:

$$x(t) = 2 - \cos(4t),\ \ y(t) = 2 - \sin(4t),$$

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:

$$\begin{aligned}\rho(t) =& [2 - \cos(4t) - 2\sqrt{2}t]^2 + [2 - \sin(4t) - 2\sqrt{2}t]^2,\\ \rho'(t) =& 16(\cos(4t) - \sin(4t))(\sqrt{2}t - 1) + 32t+\\ &+ 4\sqrt{2}(\cos(4t) + \sin(4t)) - 16\sqrt{2}.\end{aligned}$$

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. □

5.98. 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

$$x(t) = v_0 t \cos\varphi + c_1, \qquad y(t) = v_0 t \sin\varphi - \tfrac{1}{2} g t^2 + c_2$$

5.39. Theorem. *Let us suppose that f and g are functions differentiable on some neighborhood of a point $x_0 \in \mathbb{R}$, yet not necessarily at the point x_0 itself. Moreover, let the limits*

$$\lim_{x\to x_0} f(x) = 0, \qquad \lim_{x\to x_0} g(x) = 0$$

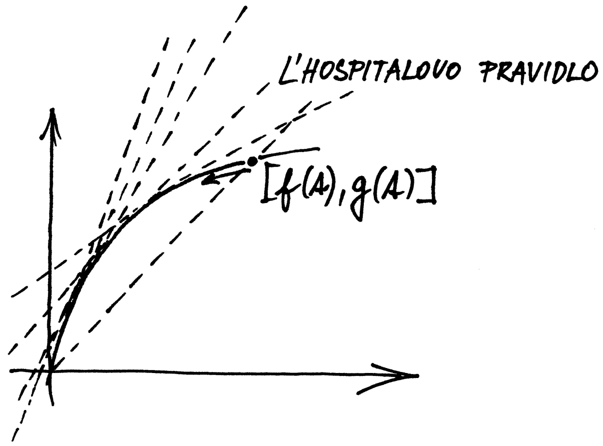
exist. If the limit

$$\lim_{x\to x_0} \frac{f'(x)}{g'(x)}$$

exists, then the limit

$$\lim_{x\to x_0} \frac{f(x)}{g(x)}$$

exists as well and these two limits are equal.



Proof. Without loss of generality, we can assume that both the functions f and g take zero at the point x_0. Again, we can illustrate the statement by a picture. Let us consider the points $[g(x), f(x)] \in \mathbb{R}^2$ parametrized by the variable x. 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, we know that the quotient of the derivatives corresponds to the slope of the secant line at the given point. Thus we want to derive 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 make use of the mean value theorem in a parametric form. First of all, let us realize that the existence of the expression $f'(x)/g'(x)$ on some neighborhood of the point x_0 (excluding x_0 itself) is implicitly assumed; thus especially for points c sufficiently close to x_0, we will have $g'(c) \neq 0$.[4] Thanks to the mean value theorem, we have now that

$$\lim_{x\to x_0} \frac{f(x)}{g(x)} = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{g(x) - g(x_0)} = \lim_{x\to x_0} \frac{f'(c_x)}{g'(c_x)},$$

where c_x is a number lying between x_0 and x, dependent on x. From 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

[4] 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.'

for $t \in (0, t_0)$ and $c_1, c_2 \in \mathbb{R}$. From the initial conditions

$$\lim_{t\to 0+} x(t) = x(0) = 0, \qquad \lim_{t\to 0+} y(t) = y(0) = 0,$$

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

$$t_0 = \frac{l}{v_0\cos\varphi}, \tag{5.1}$$

and thus we get only one equation

$$h = l\tan\varphi - \frac{gl^2}{2v_0^2\cos^2\varphi}, \tag{5.2}$$

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 (‖5.2‖). Since

$$\tfrac{1}{\cos^2\varphi} = \tfrac{\cos^2\varphi+\sin^2\varphi}{\cos^2\varphi} = 1 + \tan^2\varphi, \qquad \varphi \in \left(0, \tfrac{\pi}{2}\right),$$

the equation (‖5.2‖) can be written in the form

$$h - l\tan\varphi + \tfrac{gl^2}{2v_0^2}\left(1 + \tan^2\varphi\right) = 0,$$

i. e.

$$\tan^2\varphi - \tfrac{2v_0^2}{gl}\tan\varphi + \tfrac{2hv_0^2}{gl^2} + 1 = 0.$$

>From the last equation (quadratic equation in $p = \tan\varphi$), it follows that

$$\tan\varphi = \frac{\frac{2v_0^2}{gl} \pm \sqrt{\frac{4v_0^4}{g^2l^2} - 4\left(\frac{2hv_0^2}{gl^2} + 1\right)}}{2},$$

i. e.

$$\tan\varphi = \frac{v_0^2}{gl} \pm \frac{\sqrt{v_0^4 - 2hv_0^2 g - g^2l^2}}{gl}. \tag{5.3}$$

Therefore, the angle φ satisfying (‖5.2‖) exists if and only if

$$v_0^4 - 2gh\, v_0^2 - g^2l^2 \ge 0.$$

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

$$v_0 = \sqrt{g\left[h + \sqrt{h^2+l^2}\right]} \tag{5.4}$$

$f'(x)/g'(x)$. Especially, we can substitute any sequence c_{x_n} for $x_n \to x_0$, and thus the limit

$$\lim_{x\to x_0} \frac{f'(c_x)}{g'(c_x)}$$

will exist, and the last two limits will be equal. Thus we have proved that the wanted limit exists and also has the same value. □

>From the proof of the theorem, it is apparent that it holds for one-sided limits as well.

5.40. Corollaries. L'Hospital's rule can easily be extended for limits at the improper points ±∞ and for the case of infinite values of the limits. If, for instance, we have

$$\lim_{x\to\infty} f(x) = 0, \quad \lim_{x\to\infty} g(x) = 0,$$

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, we get

$$\lim_{x\to 0_+} \frac{(f(1/x))'}{(g(1/x))'} = \lim_{x\to 0_+} \frac{f'(1/x)(-1/x^2)}{g'(1/x)(-1/x^2)}$$
$$= \lim_{x\to 0_+} \frac{f'(1/x)}{g'(1/x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

Applying the previous theorem, we get that the limit

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to 0_+} \frac{f(1/x)}{g(1/x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}$$

will exist in this case as well.

The limit calculation is even simpler in the case when

$$\lim_{x\to x_0} f(x) = \pm\infty, \quad \lim_{x\to x_0} g(x) = \pm\infty.$$

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 holds in the same form for infinite limits as well:

Theorem. *Let f and g be functions differentiable on some neighborhood of a point $x_0 \in \mathbb{R}$, yet not necessarily at the point 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

$$\lim_{x\to x_0} \frac{f(x)}{g(x)}$$

exists as well and they equal each other.

Proof. Once again, we can 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 neighborhood 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.

is then matched (see (‖5.3‖)) by

(5.5)

$$\tan\varphi = \frac{v_0^2}{gl} = \frac{h+\sqrt{h^2+l^2}}{l}, \quad \text{i. e.} \quad \varphi = \operatorname{arctg}\frac{h+\sqrt{h^2+l^2}}{l}.$$

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 (‖5.4‖) and φ from (‖5.5‖).

Let us remind that we have (see (‖5.1‖), (‖5.2‖))

$$t_0 = \tfrac{l}{v_0\cos\varphi}, \qquad v_0^2 = \tfrac{gl^2}{2(l\tan\varphi - h)\cos^2\varphi}.$$

Using this, from

$$y'(t_0) = \lim_{t\to t_0-} y'(t) = v_0\sin\varphi - gt_0 < 0$$

we get

$$\tfrac{gl^2}{2(l\tan\varphi-h)\cos^2\varphi} = v_0^2 < v_0\cdot\tfrac{gt_0}{\sin\varphi} = \tfrac{gl}{\sin\varphi\ \cos\varphi},$$

i. e. the equality

$$l\ \sin\varphi\ \cos\varphi < 2(l\tan\varphi - h)\cos^2\varphi,$$

from which we can easily see that

$$\tfrac{2h}{l} < \tan\varphi.$$

By confrontation with (‖5.5‖), we get that the last inequality really holds because

$$\tan\varphi = \tfrac{h+\sqrt{h^2+l^2}}{l} > \tfrac{h+\sqrt{h^2}}{l} = \tfrac{2h}{l}.$$

Thus we have shown that for the initial speed from (‖5.4‖), 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

$$\varphi = \operatorname{arctg}\tfrac{v_0^2}{9.806\,65\cdot 4.225} \doteq 0.907\ \text{rad} \approx 52\,^\circ.$$

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

$$\varphi = \tfrac{\pi}{2} - \beta = \tfrac{\pi}{4} + \tfrac{\gamma}{2} = \tfrac{1}{2}\left(\tfrac{\pi}{2}+\gamma\right) = \tfrac{1}{2}\left(\tfrac{\pi}{2} + \operatorname{arctg}\tfrac{h}{l}\right).$$

Using the mean value theorem, we can replace the fraction in the middle with the quotient of the derivatives at an appropriate point c between x and y, and 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. Having fixed y and x approaching x_0, the former fraction apparently converges to one. If we simultaneously move y towards x_0, the latter fraction will get arbitrarily close to the limit value of the quotient of the derivatives. □

5.41. Use cases. Making suitable modifications of the examined expressions, one can also apply l'Hospital's rule on forms of the types $\infty - \infty$, 1^∞, $0\cdot\infty$, and so on. One often simply rearranges the expressions or uses some smooth function, for instance the exponential one.

For an illustration of such a procedure, we will show the connection between the arithmetic and geometric means of n nonnegative values x_i. The *arithmetic mean*

$$M^1(x_1,\dots,x_n) = \frac{x_1+\cdots+x_n}{n}$$

is a special case of the so-called *power mean with exponent r*, also known as generalized mean:

$$M^r(x_1,\dots,x_n) = \left(\frac{x_1^r+\cdots+x_n^r}{n}\right)^{\frac{1}{r}}.$$

The special value M^{-1} is called *harmonic mean*. Now, let us calculate the limit value of M^r for r approaching zero. For this purpose, we will determine the limit by l'Hospital's rule (it it an expression of the form 0/0 and we differentiate with respect to r, while x_i are constant parameters).

The following calculation, in which we apply the chain rule and our knowledge of the derivative of the power function, must be read from the back. Existence of the last limit implies the existence of the last-but-one, and so on.

$$\begin{aligned}\lim_{r\to 0}\ln(M^r(x_1,\dots,x_n)) &= \lim_{r\to 0}\frac{\ln(\frac{1}{n}(x_1^r+\cdots+x_n^r))}{r}\\ &= \lim_{r\to 0}\frac{\frac{x_1^r\ln x_1+\cdots+x_n^r\ln x_n}{n}}{\frac{x_1^r+\cdots+x_n^r}{n}}\\ &= \frac{\ln x_1+\cdots+\ln x_n}{n} = \ln\sqrt[n]{x_1\cdot\dots\cdot x_n}.\end{aligned}$$

Hence we can immediately see that

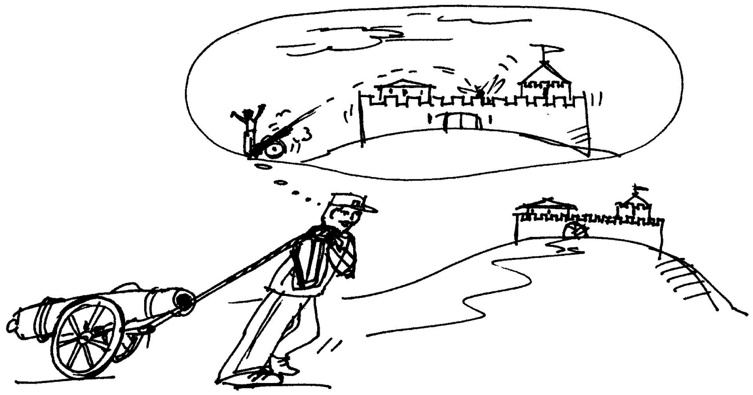
$$\lim_{r\to 0} M^r(x_1,\dots,x_n) = \sqrt[n]{x_1\dots x_n},$$

which is a value known as *geometric mean*.

4. Power series

5.42. The calculation of e^x. Besides addition and multiplication, we can also manipulate with limits of sequences. Thus it might be a good idea to approximate non-polynomial functions by sequences of values that can be calculated.

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 $\varphi = 45\,°$). 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. □

5.99. 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

$$x(t) = v_0 t \cos\varphi, \quad y(t) = v_0 t \sin\varphi - \tfrac{1}{2} g t^2, \qquad t \in (0, t_0),$$

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

Having a looking at the function e^x, we actually 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. If we apply the interest rate per year once a month, once a day, once an hour, and so on, we will get the following values for the yield x of the deposit:

$$\left(1 + \frac{x}{12}\right)^{12}, \quad \left(1 + \frac{x}{365}\right)^{365}, \quad \left(1 + \frac{x}{8760}\right)^{8760}, \quad \dots$$

Therefore, we could deem 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 one.

Let us, in detail, examine the sequence of numbers

$$a_n = \left(1 + \frac{1}{n}\right)^n.$$

The so-called *Bernoulli's inequality* will come in handy:

Lemma. *For every real number $b \ge -1$, $b \ne 0$, and a natural number $n \ge 2$, it is true that $(1 + b)^n > 1 + nb$.*

Proof. For $n = 2$, we get

$$(1 + b)^2 = 1 + 2b + b^2 > 1 + 2b.$$

From now on, we proceed by induction on n, supposing $b > -1$. Let us assume that the proposition holds for some $k \ge 2$ and let us calculate:

$$\begin{aligned} (1 + b)^{k+1} &= (1 + b)^k (1 + b) > (1 + kb)(1 + b) \\ &= 1 + (k + 1)b + kb^2 > 1 + (k + 1)b \end{aligned}$$

The statement is, of course, true for $b = -1$ as well. □

Now we can bound the quotient of adjacent terms a_n of out sequence

$$\begin{aligned} \frac{a_n}{a_{n-1}} &= \frac{(1 + \frac{1}{n})^n}{(1 + \frac{1}{n-1})^{n-1}} = \frac{(n^2 - 1)^n n}{n^{2n}(n - 1)} = \left(1 - \frac{1}{n^2}\right)^n \frac{n}{n - 1} \\ &> (1 - \frac{1}{n})\frac{n}{n - 1} = 1. \end{aligned}$$

Thus we have proved that our sequence is indeed increasing.

The following, very similar, calculation (applying Bernoulli's inequality once again) 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. Surely $b_n > a_n$.

$$\begin{aligned} \frac{b_n}{b_{n+1}} &= \frac{n}{n + 1}\left(\frac{\frac{n+1}{n}}{\frac{n+2}{n+1}}\right)^{n+2} = \frac{n}{n + 1}\left(\frac{n^2 + 2n + 1}{n^2 + 2n}\right)^{n+2} \\ &= \frac{n}{n + 1}\left(1 + \frac{1}{n(n + 2)}\right)^{n+2} \\ &\ge \frac{n}{n + 1}\left(1 + \frac{n + 2}{n(n + 2)}\right)^{n+2} = 1. \end{aligned}$$

Thus the sequence a_n is increasing and bounded from above, so the set of its terms has a supremum which equals the limit of the

$$R = v_0 t_0 \cos\varphi, \quad -h = v_0 t_0 \sin\varphi - \tfrac{1}{2}\, g t_0^2.$$

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

$$-h = R\tan\varphi - \frac{gR^2}{2v_0^2\cos^2\varphi}, \tag{5.6}$$

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 (‖5.6‖) (it is determined by the equation (‖5.6‖)). Thus, this function is given implicitly. The equation (‖5.6‖) can be written as (R is substituted by $R(\varphi)$)

$$R(\varphi)\tan\varphi \cdot 2v_0^2\cos^2\varphi - gR^2(\varphi) + h \cdot 2v_0^2\cos^2\varphi = 0.$$

Using the relation

$$2\tan\varphi\cos^2\varphi = \sin 2\varphi,$$

we can transform (‖5.6‖) into the form

$$R(\varphi)v_0^2\sin 2\varphi - gR^2(\varphi) + 2hv_0^2\cos^2\varphi = 0. \tag{5.7}$$

Differentiating with respect to φ now gives

$$R'(\varphi)v_0^2\sin 2\varphi + 2R(\varphi)v_0^2\cos 2\varphi - 2gR(\varphi)R'(\varphi) -$$
$$2hv_0^2\,(2\cos\varphi\sin\varphi) = 0,$$

i. e.

$$R'(\varphi)\left[v_0^2\sin 2\varphi - 2gR(\varphi)\right] = -2R(\varphi)v_0^2\cos 2\varphi + 2hv_0^2\sin 2\varphi.$$

Thus we have calculated that

$$R'(\varphi) = \tfrac{2v_0^2[h\sin 2\varphi - R(\varphi)\cos 2\varphi]}{v_0^2\sin 2\varphi - 2gR(\varphi)}, \qquad \varphi \in \left(0, \tfrac{\pi}{2}\right).$$

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 (‖5.6‖), obtaining

$$-h = \tfrac{v_0^2\sin\varphi\cos\varphi}{g}\tan\varphi - \tfrac{gv_0^4\sin^2\varphi\cos^2\varphi}{2g^2v_0^2\cos^2\varphi}.$$

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).

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

$$h\sin 2\varphi = R(\varphi)\cos 2\varphi, \qquad \text{i. e.} \qquad R(\varphi) = h\tan 2\varphi.$$

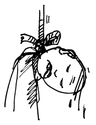
Since the function R must have a maximum on the interval $(0, \pi/2)$ (according to the physical meaning of the problem, for $\varphi \to 0+$ and sequence. At the same time, we can see that this value is also the limit of the decreasing sequence b_n because

$$\lim_{n\to\infty} b_n = \lim_{n\to\infty}(1 + \frac{1}{n})a_n = \lim_{n\to\infty} a_n.$$

This limit determines one of the most important numbers in mathematics (besides the numbers 0, 1, and π), the so-called *Euler's number* e. We thus have

$$\mathrm{e} = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n.$$

5.43. Power series for e^x. The exponential function has been defined as the only continuous function satisfying $f(1) = \mathrm{e}$ 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



$$\mathrm{e}^x = \lim_{n\to\infty}(a_n)^x.$$

For the sake of simplicity, let us fix a positive number x. If we replace n with n/x in the values a_n from the previous paragraph, we again arrive at the same limit. (Think this out in detail!) Hence

$$\mathrm{e} = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^{\frac{n}{x}}, \quad \mathrm{e}^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n.$$

Let us denote the n-th term of this sequence as $u_n(x) = (1 + x/n)^n$ and express it by the binomial theorem:

$$\begin{aligned} u_n(x) &= 1 + n\frac{x}{n} + \frac{n(n-1)x^2}{2!n^2} + \cdots + \frac{n!x^n}{n!n^n} \\ &= 1 + x + \frac{x^2}{2!}\left(1 - \frac{1}{n}\right) \\ &\quad + \frac{x^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \dots \\ &\quad + \frac{x^n}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\dots\left(1 - \frac{n-1}{n}\right). \end{aligned} \tag{5.10}$$

Since all the expressions in the parentheses are less than one, we also get that

$$u_n(x) < v_n(x) = \sum_{j=0}^{n}\frac{1}{j!}x^j.$$

Let us have a look at the formal infinite sum

$$\sum_{j=0}^{\infty} c_j = \sum_{j=0}^{\infty}\frac{1}{j!}x^j. \tag{5.11}$$

We can see that $v_n(x)$ is just the sum of the first n terms of this formal infinite expression.

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 $j \geq N$. However, so great indeces j satisfy $c_{j+1} < \frac{1}{2}c_j < 2^{-(j-N+1)}\,c_N$. This means that the parial sums of the first n terms of our formal sum are bounded from above by the sums

$$v_n < \sum_{j=0}^{N-1}\frac{1}{j!}x^j + \frac{1}{N!}x^N\sum_{j=0}^{n-N}\frac{1}{2^j}.$$

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

$$R\left(\varphi\right) = h \tan 2\varphi. \tag{5.8}$$

Let us thus substitute (‖5.8‖) into (‖5.7‖). We obtain

$$h \tan 2\varphi \, v_0^2 \sin 2\varphi - gh^2 \tan^2 2\varphi + 2hv_0^2 \cos^2 \varphi = 0.$$

This equation can be transformed to

$$\tan 2\varphi \, v_0^2 \sin 2\varphi + 2v_0^2 \cos^2 \varphi = gh \tan^2 2\varphi,$$

$$v_0^2 \tfrac{\sin^2 2\varphi}{\cos 2\varphi} + v_0^2 \left(\cos 2\varphi + 1\right) = gh \tfrac{\sin^2 2\varphi}{\cos^2 2\varphi},$$

$$v_0^2 \sin^2 2\varphi + v_0^2 \cos^2 2\varphi + v_0^2 \cos 2\varphi = gh \tfrac{\sin^2 2\varphi}{\cos 2\varphi},$$

$$v_0^2 + v_0^2 \cos 2\varphi = gh \tfrac{1-\cos^2 2\varphi}{\cos 2\varphi},$$

$$v_0^2 \left(1 + \cos 2\varphi\right) = gh \tfrac{(1-\cos 2\varphi)(1+\cos 2\varphi)}{\cos 2\varphi},$$

$$v_0^2 \cos 2\varphi = gh \left(1 - \cos 2\varphi\right),$$

$$\left(v_0^2 + gh\right) \cos 2\varphi = gh,$$

$$\cos 2\varphi = \tfrac{gh}{v_0^2+gh}.$$

However, by this we have uniquely determined the point

$$\varphi_0 = \tfrac{1}{2} \arccos \tfrac{gh}{v_0^2+gh},$$

at which R is highest. Since

$$\sin 2\varphi_0 = \sqrt{1 - \cos^2 2\varphi_0} = \sqrt{1 - \tfrac{g^2h^2}{\left(v_0^2+gh\right)^2}} = \tfrac{\sqrt{v_0^4+2ghv_0^2}}{v_0^2+gh},$$

the function value

$$R\left(\varphi_0\right) = h \tan 2\varphi_0 = h \, \frac{\frac{\sqrt{v_0^4+2ghv_0^2}}{v_0^2+gh}}{\frac{gh}{v_0^2+gh}} = \frac{\sqrt{v_0^4+2ghv_0^2}}{g} = \frac{v_0}{g} \sqrt{v_0^2 + 2gh}.$$

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 $g = 9.806\,65 \text{ m} \cdot \text{s}^{-2}$). Then the javelin can fly up to the distance

$$R\left(\varphi_0\right) = \tfrac{27.778}{9.806\,65} \sqrt{27.778^2 + 2 \cdot 9.806\,65 \cdot 1.8} \text{ m} \doteq 80.46 \text{ m}.$$

This distance was achieved for

$$\varphi_0 = \tfrac{1}{2} \arccos \tfrac{9.806\,65 \cdot 1.8}{27.778^2 + 9.806\,65 \cdot 1.8} \doteq 0.774\,2 \text{ rad} \approx 44.36\,°.$$

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 %. The original record (with "correctly balanced" javelin) was 80.00 m.

Since for every q, it holds that $(1-q)(1+q+\cdots+q^k) = 1-q^{k+1}$, we can also bound the values v_n:

$$v_n < \sum_{j=0}^{N-1} \frac{1}{j!} x^j + \frac{2}{N!} x^N \left(1 - 2^{-n+N-1}\right)$$

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 exists as well.

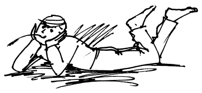
Now, let us examine the sequence of numbers u_n, whose limit is e^x. We will consider $n > N$ for some fixed N (very great) and choose a fixed number $k < N$ (quite small). Then for sufficiently large N, we can approximate the sum of the first k terms in the expression of u_N in (5.10) by v_k with arbitrary precision. Since this part of the sum of u_N is strictly less than u_N itself, the sequence u_n must converge to the same limit as the sequence v_n. Thus we have proved

The power series for e^x

Theorem. *The exponential function is, for every number $x \in \mathbb{R}$, expressed as the limit of the partial sums in the expression*

$$\mathrm{e}^x = 1 + x + \frac{1}{2!} x^2 + \cdots + \frac{1}{n!} x^n + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!} x^n .$$

5.44. Number series. When deriving the previous important theorems about the function e^x, we have accidentally worked with several extraordinarily useful concepts and tools. Now, we will formulate them in general:



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 $s_k = \sum_{n=0}^{k} a_n$, and we say that the series converges and equals s iff 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, we say that the series *diverges* to ∞ or $-\infty$. If the limit of the partial sums does not exist, we sometimes say that the series *oscillates*.

For the sequence of partial sums s_n to be converging, it is necessary and sufficient that it is a Cauchy sequence, i. e.

$$|s_m - s_n| = |a_{n+1} + \cdots + a_m|$$

must be arbitrarily small for sufficiently great $m > n$. 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 $\sum_{k=0}^{\infty} a_n$.

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

$$\varphi_0 = \lim_{h\to 0+} \tfrac{1}{2} \arccos \tfrac{gh}{v_0^2+gh} = \tfrac{1}{2} \arccos 0 = \tfrac{\pi}{4} \text{ rad} = 45\,^\circ$$

at which the ball falls at the greatest distance

$$R\,(\varphi_0) = \lim_{h\to 0+} \tfrac{v_0}{g} \sqrt{v_0^2 + 2gh} = \tfrac{v_0^2}{g}.$$

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. □

Further miscellaneous problems concerning extrema of functions of a real variable can be found at 320

G. L'Hospital's rule

5.100. Verify that the limit

(a)

$$\lim_{x\to 0} \frac{\sin(2x) - 2\sin x}{2\mathrm{e}^x - x^2 - 2x - 2} \quad \text{is of the type } \frac{0}{0};$$

(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)

$$\lim_{x\to 1+} (\ln(x-1)\cdot \ln x) \quad \text{is of the type } 0\cdot\infty;$$

(e)

$$\lim_{x\to 0+} (\cot x)^{\frac{1}{\ln x}} \quad \text{is of the type } \infty^0;$$

(f)

$$\lim_{x\to 0} \left(\frac{\sin x}{x} \right)^{\frac{1}{x^2}} \quad \text{is of the type } 1^\infty;$$

Absolutely convergent series

We say that a series $\sum_{k=0}^{\infty} a_n$ is *absolutely convergent* iff the series $\sum_{n=0}^{\infty} |a_n|$ converges.

The absolute convergence has been introduced because it can often be much easily verified. Moreover, the following theorem shows that all simple algebraic operations behave "very well" in the case of series that converge absolutely.

5.45. 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

$$S\cdot T = \left(\sum_{n=0}^{\infty} a_n\right)\cdot\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_{n-k} b_k\right).$$

Proof. Both the first and the second statements are a straightforward consequence of the corresponding properties of limits. The third statement requires our attention. Let us write

$$c_n = \sum_{k=0}^{n} a_{n-k} b_k.$$

From the assumptions and the rule for the limit of a product, we get

$$\left(\sum_{n=0}^{k} a_n\right)\cdot\left(\sum_{n=0}^{k} b_n\right) \to \left(\sum_{n=0}^{\infty} a_n\right)\cdot\left(\sum_{n=0}^{\infty} b_n\right).$$

Thus it suffices to prove that

$$0 = \lim_{k\to\infty} \left(\left(\sum_{n=0}^{k} a_n\right)\cdot\left(\sum_{n=0}^{k} b_n\right) - \sum_{n=0}^{k} c_k \right).$$

Let us confront the expressions

$$\left(\sum_{n=0}^{k} a_n\right)\cdot\left(\sum_{n=0}^{k} b_n\right) = \sum_{0\le i,j\le k} a_i b_j,$$

$$c_n = \sum_{\substack{i+j=n\\ 0\le i,j\le k}} a_i b_j, \quad \sum_{n=0}^{k} c_n = \sum_{\substack{i+j\le k\\ 0\le i,j\le k}} a_i b_j.$$

Thus we get the bound

$$\left| \left(\sum_{n=0}^{k} a_n\right)\cdot\left(\sum_{n=0}^{k} b_n\right) - \sum_{n=0}^{k} c_n \right| = \left| \sum_{\substack{i+j>k\\ 0\le i,j\le k}} a_i b_j \right| \le \sum_{\substack{i+j>k\\ 0\le i,j\le k}} |a_i b_j|.$$

To bound the last expression, we can use a simple trick: If the sum of the indeces is to be greater than k, then at least one of them must be greater than $k/2$. Surely we will not lower the expression if we

(g)
$$\lim_{x\to 1-}\left(\cos\frac{\pi x}{2}\right)^{\ln x} \qquad \text{is of the type } 0^0.$$
Then calculate it using l'Hospital's rule.

Solution. We can immediately assert that

(a)
$$\lim_{x\to 0}\left(\sin(2x)-2\sin x\right)=0-0=0,$$
$$\lim_{x\to 0}\left(2\mathrm{e}^x-x^2-2x-2\right)=2-0-0-2=0;$$

(b)
$$\lim_{x\to 0+}\ln x=-\infty, \qquad \lim_{x\to 0+}\cot x=+\infty;$$

(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
$$\lim_{x\to 0}\frac{\sin(2x)-2\sin x}{2\mathrm{e}^x-x^2-2x-2}$$
into the limit
$$\lim_{x\to 0}\frac{2\cos(2x)-2\cos x}{2\mathrm{e}^x-2x-2},$$
which is of the type $0/0$. Two more applications of the rule lead to
$$\lim_{x\to 0}\frac{-4\sin(2x)+2\sin x}{2\mathrm{e}^x-2}$$
and (the above limit is also of the type $0/0$)
$$\lim_{x\to 0}\frac{-8\cos(2x)+2\cos x}{2\mathrm{e}^x}=\frac{-8+2}{2}=-3.$$
Altogether, we have (returning to the original limit)
$$\lim_{x\to 0}\frac{\sin(2x)-2\sin x}{2\mathrm{e}^x-x^2-2x-2}=-3.$$
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. add more terms into it, i. e. we will take all as in the product and remove only those whose indeces are both at most $k/2$.
$$\sum_{\substack{i+j>k\\ 0\le i,j\le k}}|a_ib_j| \le \sum_{0\le i,j\le k}|a_ib_j| - \sum_{0\le i,j\le k/2}|a_ib_j|.$$
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 following theorem states some conditions that will help us verify the convergence of series.

5.46. Theorem. *Let $S=\sum_{n=0}^{\infty}a_n$ be an infinite series of real or complex numbers.*

(1) If the series S converges, then $\lim_{n\to\infty}a_n=0$.

(2) Let us 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 $|q|>1$. We know nothing in the case of $|q|=1$: the series may or may not converge.

(3) If the limit
$$\lim_{n\to\infty}\sqrt[n]{|a_n|}=q$$
exists, then the series converges absolutely for $q<1$ while it does not converge for $q>1$. Again, in the case of $q=1$, the series may or may not converge.

Proof. (1) We know that 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 sufficiently small number $\varepsilon>0$, there are infinitely many terms a_k with $|a_k|>\varepsilon$. So there must be either infinitely many positive terms of infinitely many negative terms among them. But then, adding any one of them into the partial sum, we get that 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 converging, either.

(2) Since we want to prove the absolute convergence, we can assume straightaway that the terms of the series are real numbers $a_i>0$. The proof was given for the special value of $q=1/2$ when deriving the value of e^x using series. Now, let us consider $q<r<1$ for a real number r. From the existence of the limit of the quotients, we can deduce that for every j greater than a sufficiently large N, it holds that
$$a_{j+1}<r\cdot a_j\le r^{(j-N+1)}\,a_N.$$
But this means that the partial sums s_n are, for large $n>N$, bounded from above by the sums
$$s_n<\sum_{j=0}^{N}a_j+a_N\sum_{j=0}^{n-N}r^j=\sum_{j=0}^{N}a_j+\frac{1-r^{n-N+1}}{1-r}.$$
Since $0<r<1$, the set of all partial sums is an increasing sequence bounded from above, and thus its limits is its supremum.

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

$$\lim_{x\to 0+} -\sin x = 0, \qquad \lim_{x\to 0+} \frac{\sin x}{x} = 1,$$

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:

$$\lim_{x\to 1+} \frac{-x \ln^2 x}{x-1} = \lim_{x\to 1+} \frac{-\ln^2 x - 2x \ln x \cdot \frac{1}{x}}{1} = -\frac{0+0}{1} = 0.$$

In the case $q > r > 1$, we will use a similar technique. However, this time, from the existence of the limit of the quotients, we now deduce that

$$a_{j+1} > r \cdot a_j \geq r^{(j-N+1)} \, a_N > 0.$$

However, this means that the absolute values of the particular terms of the series do not converge to zero, and thus the series cannot be converging, by the already proved part of the theorem.

(3) The proof is quite 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 us $|a_n| < r^n$, so we, once again, are comparing this to a geometric series. Thus the proof can be finished in the same way as in the case of the ratio test. □

In the proof of both the second and the third statement, we have used a weaker assumption than the existence of the limit. We only wanted to know that the examined sequences of non-negative terms are, from a given index on, either all greater 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 *limes superior* of the sequence and denoted by

$$\limsup_{n\to\infty} b_n.$$

The advantage is that limes superior 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.

5.47. 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 "pointwise", thereby obtaining the concept of the *series of functions*

$$S(x) = \sum_{n=0}^{\infty} f_n(x).$$

Convergence of power series

A *power series* is given by an expression

$$S(x) = \sum_{n=0}^{\infty} a_n x^n.$$

The cases (e), (f), (g). Since

$$\lim_{x\to 0+}(\cot x)^{\frac{1}{\ln x}} = \mathrm{e}^{\lim\limits_{x\to 0+}\frac{\ln(\cot x)}{\ln x}};$$

$$\lim_{x\to 0}\left(\frac{\sin x}{x}\right)^{\frac{1}{x^2}} = \mathrm{e}^{\lim\limits_{x\to 0}\frac{\ln\frac{\sin x}{x}}{x^2}};$$

$$\lim_{x\to 1-}\left(\cos\frac{\pi x}{2}\right)^{\ln x} = \mathrm{e}^{\lim\limits_{x\to 1-}\left(\ln x\cdot\ln(\cos\frac{\pi x}{2})\right)},$$

it suffices to calculate the limits given in the argument of the exponential function. By l'Hospital's rule and simple rearrangements, we get

$$\lim_{x\to 0+}\frac{\ln(\cot x)}{\ln x}\left[\text{type }\frac{+\infty}{-\infty}\right] = 1\lim_{x\to 0+}\frac{\frac{1}{\cot x}\cdot\frac{-1}{\sin^2 x}}{\frac{1}{x}} =$$

$$\lim_{x\to 0+}\frac{-x}{\cos x\cdot\sin x}\left[\text{type }\frac{0}{0}\right] = \lim_{x\to 0+}\frac{-1}{\cos^2 x-\sin^2 x} = \frac{-1}{1-0} = -1;$$

$$\lim_{x\to 0}\frac{\ln\frac{\sin x}{x}}{x^2}\left[\text{type }\frac{0}{0}\right] = \lim_{x\to 0}\frac{\frac{x}{\sin x}\cdot\frac{x\cos x-\sin x}{x^2}}{2x} = \lim_{x\to 0}\frac{x\cos x-\sin x}{2x^2\sin x}$$

$$\left[\text{type }\frac{0}{0}\right] = \lim_{x\to 0}\frac{\cos x-x\sin x-\cos x}{4x\sin x+2x^2\cos x}$$

$$= \lim_{x\to 0}\frac{-\sin x}{4\sin x+2x\cos x}\left[\text{type }\frac{0}{0}\right]$$

$$= \lim_{x\to 0}\frac{-\cos x}{4\cos x+2\cos x-2x\sin x} = \frac{-1}{4+2-0} = -\frac{1}{6},$$

hence

$$\lim_{x\to 0+}(\cot x)^{\frac{1}{\ln x}} = \mathrm{e}^{-1} = \frac{1}{\mathrm{e}};$$

$$\lim_{x\to 0}\left(\frac{\sin x}{x}\right)^{\frac{1}{x^2}} = \mathrm{e}^{-\frac{1}{6}} = \frac{1}{\sqrt[6]{\mathrm{e}}}.$$

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

$$\lim_{x\to 1-}\left(x\sin\frac{\pi x}{2}\right)\cdot\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}} = 1\cdot\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}}.$$

We say that $S(x)$ has *radius of convergence* $\rho\geq 0$ iff $S(x)$ converges for every x satisfying $|x|<\rho$ and does not converge for $|x|>\rho$.

5.48. Properties of power series. Although a significant part of the proof of the following theorem will have to be postponed until the end of the following chapter, we will formulate the basic properties of the power series right now:

ABSOLUTE CONVERGENCE AND DIFFERENTIATION

Theorem. *Let $S(x)=\sum_{n=0}^{\infty}a_nx^n$ be a power series and let the limit*

$$r=\lim_{n\to\infty}\sqrt[n]{|a_n|}$$

exist. Then the radius of convergence of the series S equals $\rho=r^{-1}$.

The power series $S(x)$ converges absolutely on the whole interval of convergence and is continuous on it (including the marginal points, supposing it is convergent there). Moreover, the derivative exists on this interval, and

$$S'(x)=\sum_{n=1}^{\infty}na_nx^{n-1}.$$

PROOF. To verify the absolute convergence of the series, we can use the root test from theorem 5.46(3), for every value of x. We calculate

$$\lim_{n\to\infty}\sqrt[n]{|a_nx^n|}=rx,$$

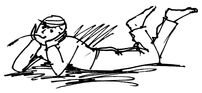
and the series converges absolutely, or does not converge if this limit is different from 1. Hence it follows that it indeed converges for $|x|<\rho$ and diverges for $|x|>\rho$.

The statements about the continuity and the derivatives will be proved later in a more general context, see 6.43–6.45. □

Let us also notice that, when proving the convergence, we can use a stronger form of the root test, and so the radius r of convergence can, for every power series, be described explicitly by

$$r^{-1}=\limsup_{n\to\infty}\sqrt[n]{|a_n|}.$$

5.49. Notes. If the coefficients of the series increase rapidly, i. e. $a_n=n^n$, then $r=\infty$, i. e. the radius of convergence is zero. Indeed, such a series converges at a single point, namely $x=0$.



Now we will have a look at some examples of convergence of power series (including the marginal points of the corresponding interval): Let us consider

$$S(x)=\sum_{n=0}^{\infty}x^n,\quad T(x)=\sum_{n=1}^{\infty}\frac{1}{n}x^n.$$

The former case is a *geometric series*, which we have already met. 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 apparently divergent. For $x=-1$, we get the series $1-1+1-\dots$, whose partial sums do not have a limit, i. e. the series oscillates.

Only now we apply l'Hospital's rule for

$$\lim_{x\to 1-}\frac{\ln^2 x}{\cos\frac{\pi x}{2}} = \left[\text{type } \frac{0}{0}\right] = \lim_{x\to 1-}\frac{2\ln x\cdot\frac{1}{x}}{\left(-\frac{\pi}{2}\right)\sin\frac{\pi x}{2}} = \frac{0}{-\frac{\pi}{2}} = 0.$$

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.

$$\lim_{x\to 1-}\left(\cos\frac{\pi x}{2}\right)^{\ln x} = \mathrm{e}^0 = 1.$$

□

5.101. 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

$$\frac{x-1}{x}\le\frac{x+\sin x}{x}\le\frac{x+1}{x},$$

and by the squeeze theorem,

$$1 = \lim_{x\to\infty}\frac{x+\sin x}{x}\le\lim_{x\to\infty}\frac{x+\sin x}{x}\le\lim_{x\to\infty}\frac{x+1}{x} = 1.$$

□

5.102. Determine

$$\lim_{x\to+\infty}\frac{\ln x}{x},\qquad \lim_{x\to 0+} x\ln\frac{1}{x},\qquad \lim_{x\to 0+} x\,\mathrm{e}^{\frac{1}{x}};$$

$$\lim_{x\to 0-} x\,\mathrm{e}^{-\frac{1}{x}},\qquad \lim_{x\to 0}\frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^{100}},\qquad \lim_{x\to+\infty}(\ln x - x)\,;$$

$$\lim_{x\to+\infty}\frac{x}{x+\ln x\cdot\cos x},\qquad \lim_{x\to+\infty}\frac{\sqrt[3]{x+1}}{\sqrt[5]{x+3}},\qquad \lim_{x\to+\infty}\frac{x}{\sqrt{x^2+1}}.$$

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

$$\lim_{x\to+\infty}\frac{x^n}{\mathrm{e}^x} = 0,\quad\text{i. e.}\quad \lim_{x\to+\infty}\frac{\mathrm{e}^x}{x^n} = +\infty.$$

The squeeze theorem implies the following generalization for real numbers $a>0$:

$$\lim_{x\to+\infty}\frac{x^a}{\mathrm{e}^x} = 0,\quad\text{i. e.}\quad \lim_{x\to+\infty}\frac{\mathrm{e}^x}{x^a} = +\infty.$$

The theorem 5.46(2) shows that the radius of convergence of the latter example is 1 as well 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 $x=1$, we get the series $1+\frac{1}{2}+\frac{1}{3}+\dots$, which is divergent: By gradually 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 $1/2$), we can bound the partial sums 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 it, of course, cannot converge absolutely. This follows from a more general theorem which will be introduced in the next chapter.

5.50. Trigonometric functions. With the power series, our society of functions increased by a lot of new examples of smooth functions, i. e. functions which are arbitrarily many times differentiable on the whole of their domains. Moreover, all of these additions to out menagerie have the property (similarly to polynomials) that the formula which defines them in fact defines a function $\mathbb{C}\to\mathbb{C}$.

Indeed, our reasoning about the absolute convergence holds flawlessly for series of complex numbers as well. Therefore, the power series will be convergent when we replace x with any complex number lying inside the disc with radius r centered at the origin of the complex plane.

Let us, for a while, play with the most important example, the exponential function

$$\mathrm{e}^x = 1+x+\frac{1}{2}x^2+\dots+\frac{1}{n!}x^n+\dots.$$

This power series has infinite radius of convergence, so it defines a smooth 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. Especially, the values of the power series are completely determined on the complex domain by their values at real input values x. Therefore, the complex exponential must also satisfy the usual formulas which we have already derived for the real values x. In particular, we have

$$\mathrm{e}^{x+y} = \mathrm{e}^x\cdot\mathrm{e}^y,$$

see (5.5) and the theorem 5.45(3). Let us substitute the values $x = i\cdot t$, where $i\in\mathbb{C}$ is the imaginary unit, $t\in\mathbb{R}$ arbitrary.

$$\mathrm{e}^{it} = 1+it-\frac{1}{2}t^2-i\frac{1}{3!}t^3+\frac{1}{4!}t^4+i\frac{1}{5!}t^5-\dots$$

and apparently, the conjugate number to $z=\mathrm{e}^{it}$ is the number $\bar{z}=\mathrm{e}^{-it}$. Hence

$$|z|^2 = z\cdot\bar{z} = \mathrm{e}^{it}\cdot\mathrm{e}^{-it} = \mathrm{e}^0 = 1$$

and all the values $z=\mathrm{e}^{it}$ 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 have been described using the *trigonometric functions* $\cos\theta$ and $\sin\theta$, where θ is the corresponding angle.

Taking into account that the graphs of the functions $y = \mathrm{e}^x$ and $y = \ln x$ (the inverse function to $y = \mathrm{e}^x$) are symmetric with regard to the line $y = x$, we further see that

$$\lim_{x\to+\infty} \frac{\ln x}{x} = 0, \quad \text{i. e.} \quad \lim_{x\to+\infty} \frac{x}{\ln x} = +\infty.$$

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

$$\lim_{x\to0+} x \ln \frac{1}{x} = \lim_{y\to+\infty} \frac{\ln y}{y} = 0;$$

$$\lim_{x\to0+} x\,\mathrm{e}^{\frac{1}{x}} = \lim_{y\to+\infty} \frac{\mathrm{e}^y}{y} = +\infty.$$

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,

$$\lim_{x\to0-} x\,\mathrm{e}^{-\frac{1}{x}} = \lim_{u\to+\infty} -\frac{\mathrm{e}^u}{u} = -\infty;$$

$$\lim_{x\to0} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^{100}} = \lim_{v\to+\infty} \frac{v^{50}}{\mathrm{e}^v} = 0,$$

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.

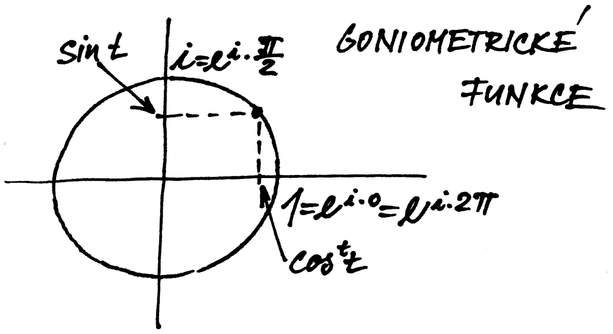
We can equally easily determine

$$\lim_{x\to+\infty} \frac{x}{x + \ln x \cdot \cos x} = \lim_{x\to+\infty} \frac{x}{x} = 1;$$

$$\lim_{x\to+\infty} \frac{\sqrt[3]{x+1}}{\sqrt[5]{x+3}} = \lim_{x\to+\infty} \frac{\sqrt[3]{x}}{\sqrt[5]{x}} = +\infty;$$

$$\lim_{x\to+\infty} \frac{x}{\sqrt{x^2+1}} = \lim_{x\to+\infty} \frac{x}{\sqrt{x^2}} = 1.$$

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


sin t
$i = e^{i\cdot\frac{\pi}{2}}$
GONIOMETRICKÉ FUNKCE
$1 = e^{i\cdot 0} = e^{i\cdot 2\pi}$
cos t

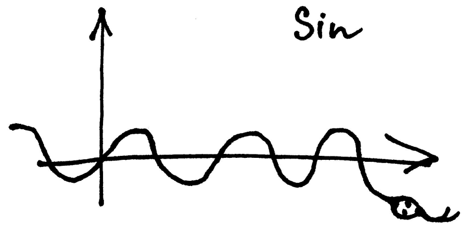
Differentiating the parametric description of the points of a circle $t \mapsto \mathrm{e}^{it}$, we get the vectors of "velocities" which will be given by the formula (if we do not believe that the power series can be differentiated term by term yet, we can instead differentiate the real part and the imaginary part separately) $t \mapsto (\mathrm{e}^{it})' = i \cdot \mathrm{e}^{it}$, so their will keep the unit size. Hence we can deem that the whole circle will be traversed when the value of the parameter reaches the length of the circle, i. e. 2π (a thorough definition of the length of a curve needs integral calculus, then we will be able to verify this statement). This procedure can be used to define the *number* π, sometimes also called Archimedes' constant or Ludolphian number [5] half the length of the unit circle in the Euclidean plane $\mathbb{R}^2$.

Now, we can at least partially convince ourselves by a look at the least positive roots of the real part of the partial sums of our series, i. e. the corresponding polynomials. Already with order ten, we get the number π with accuracy of 5 decimal places.

Thus we obtain the definition of trigonometric functions in terms of the power series:

$$\cos t = \operatorname{re} \mathrm{e}^{it} = 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 - \frac{1}{6!}t^6 + \dots + (-1)^k \frac{1}{(2k)!}t^{2k} + \dots$$

$$\sin t = \operatorname{im} \mathrm{e}^{it} = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 + \dots + (-1)^k \frac{1}{(2k+1)!}t^{2k+1} + \dots$$



The following picture illustrates the convergence of the series for the cosine function. It is the graph of the corresponding polynomial of degree 68. Gradually drawing the partial sums, we can see that the approximation near zero is very good and hardly changes at all. As the order increases, the approximation gets better farther from the origin as well.

[5] This number describes the ratio of the circumference to the diameter of an (arbitrary) circle. It was known to Babylonians and Greeks as early as the 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.

forms). If we had applied it to the first problem, we would have obtained, for $x > 0$, the quotient

$$\frac{1}{1+\frac{\cos x}{x}-\ln x\cdot\sin x}=\frac{x}{x+\cos x-x\ln x\cdot\sin x},$$

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

$$\frac{\frac{2x}{2\sqrt{x^2+1}}}{1}=\frac{x}{\sqrt{x^2+1}}.$$

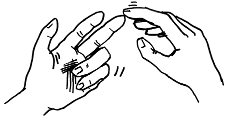
From here, we can deduce that the limit equals 1 (we are looking for a non-negative real number $a\in\mathbb{R}$ such that $a=a^{-1}$) only if we have already shown it exists at all. □

Other examples concerning calculation of limits by l'Hospital's rule can be found at page 331.

H. Infinite series

Infinite series naturally appear in a series (of problems).

5.103. 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.



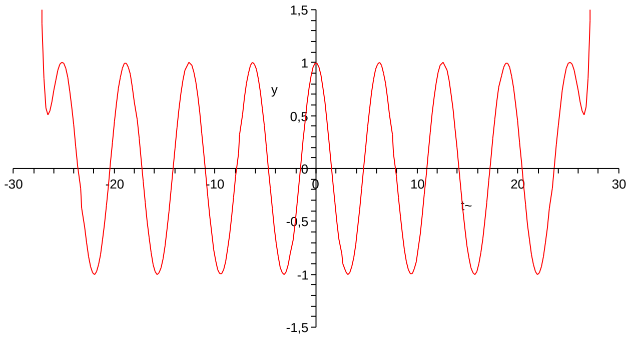
Solution. In the first step, a square having the area of $1/9$ is removed. In the second steps, eight squares (each having the area of 9^{-2}, i. e. totaling to $8\cdot 9^{-2}$) are removed. Every further iteration removes eight times more squares than in the previous steps, but the squares are nine times smaller. The sum of areas of all the removed squares is

$$\tfrac{1}{9}+\tfrac{8}{9^2}+\tfrac{8^2}{9^3}+\cdots=\sum_{n=0}^{\infty}\tfrac{8^n}{9^{n+1}}.$$

The area of the remaining figure (known as Sierpiński carpet) thus equals

$$1-\sum_{n=0}^{\infty}\tfrac{8^n}{9^{n+1}}=1-\tfrac{1}{9}\sum_{n=0}^{\infty}\left(\tfrac{8}{9}\right)^n=1-\tfrac{1}{9}\cdot\tfrac{1}{1-\frac{8}{9}}=0.$$

□



The well-known formula

$$e^{it}\,e^{-it}=\sin^2 t+\cos^2 t=1$$

follows straight from the definition. Further, from the derivative $(e^{it})'=i\,e^{it}$ we can see that

$$(\sin t)'=\cos t,\qquad (\cos t)'=-\sin t.$$

Of course, this result can also be verified by differentiating our series term by term.

Let t_0 denote the least positive number for which $e^{-it_0}=-e^{it_0}$, i. e. the first positive zero point of the function $\cos t$. According to out definition of π, we have $t_0=\frac{1}{2}\pi$.

The square of this value is $e^{i2t_0}=e^{-i2t_0}=(e^{-it_0})^2$, and so it is a zero point of the function $\sin t$. Of course, for any t, it holds that

$$e^{i(4kt_0+t)}=(e^{it_0})^{4k}\cdot e^{it}=1\cdot e^{it}\,.$$

Therefore, both trigonometric functions sin and cos are *periodic*, with period 2π. Right from our definitions, we can see that this is their prime period.

Now we can easily derive all the usual formulae connecting the trigonometric functions. We will, for illustration, introduce some of them. First, let us notice that the definition says that

$$\cos t=\frac{1}{2}(e^{it}+e^{-it})\tag{5.12}$$

$$\sin t=\frac{1}{2i}(e^{it}-e^{-it}).\tag{5.13}$$

Thus the product of these functions can be expressed as

$$\begin{aligned}\sin t\cos t&=\frac{1}{4i}(e^{it}-e^{-it})(e^{it}+e^{-it})\\&=\frac{1}{4i}(e^{i2t}-e^{-i2t})=\frac{1}{2}\sin 2t.\end{aligned}$$

Further, we can utilize our knowledge of derivatives:

$$\cos 2t=(\frac{1}{2}\sin 2t)'=(\sin t\cos t)'=\cos^2 t-\sin^2 t.$$

The properties of further 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 pictures (they are the red one and the green one on the left, and the red one and the green one on the right, respectively):

5.104. 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 from, not into, the original triangle) the middle part and remove 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

$$\lim_{n\to\infty}\left(S_0 + 3\cdot\tfrac{S_0}{9} + 4\cdot 3\cdot\tfrac{S_0}{9^2} + \cdots + 4^n\cdot 3\cdot\tfrac{S_0}{9^{n+1}}\right) =$$

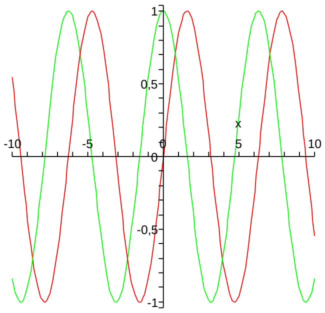
$$S_0 \lim_{n\to\infty}\left(1 + \tfrac{1}{3} + \tfrac{1}{3}\cdot\tfrac{4}{9} + \cdots + \tfrac{1}{3}\cdot\left(\tfrac{4}{9}\right)^n\right) =$$

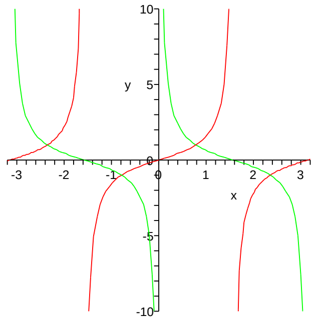
$$S_0\left[1 + \tfrac{1}{3}\lim_{n\to\infty}\left(1 + \tfrac{4}{9} + \cdots + \left(\tfrac{4}{9}\right)^n\right)\right] = S_0\left[1 + \tfrac{1}{3}\lim_{n\to\infty}\sum_{k=0}^{n}\left(\tfrac{4}{9}\right)^k\right] =$$

$$S_0\left[1 + \tfrac{1}{3}\sum_{k=0}^{\infty}\left(\tfrac{4}{9}\right)^k\right] = S_0\left[1 + \tfrac{1}{3}\cdot\tfrac{1}{1-\frac{4}{9}}\right] = \tfrac{8}{5}\,S_0.$$

The snowflake's area is thus equal to 8/5 of the area of the original triangle, i. e.

$$\tfrac{8}{5}\,S_0 = \tfrac{8}{5}\cdot\tfrac{\sqrt{3}}{4} = \tfrac{2\sqrt{3}}{5}.$$





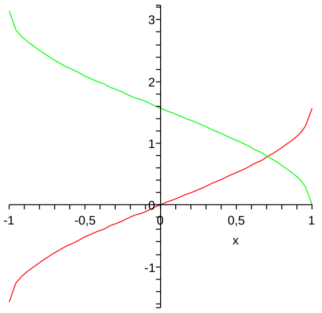
Cyclometric functions are the functions inverse to trigonometric functions. Since the trigonometric functions all have period 2π, their inverses can be defined only inside one period, and further only on the part where the given function is either increasing or decreasing. The inverse trigonometric functions are

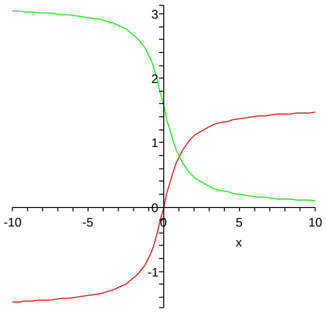
$$\arcsin = \sin^{-1}$$

with domain $[-1, 1]$ and range $[-\pi/2, \pi/2]$. Then

$$\arccos = \cos^{-1}$$

with domain $[-1, 1]$ and range $[0, \pi]$, see the left-hand picture.





The remaining functions are (displayed in the picture 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 so-called *hyperbolic functions* are also of great importance in practice, namely

$$\sinh x = \frac{1}{2}(\mathrm{e}^x - \mathrm{e}^{-x}), \qquad \cosh x = \frac{1}{2}(\mathrm{e}^x + \mathrm{e}^{-x}).$$

The name indicates that they should have something in common with a hyperbola. A straight calculation gives (the squares cancel out and only the mixed terms remain)

$$(\cosh x)^2 - (\sinh x)^2 = 2\frac{1}{2}(\mathrm{e}^x\,\mathrm{e}^{-x}) = 1.$$

The points $[\cosh t, \sinh t] \in \mathbb{R}^2$ indeed parametrically describe a hyperbola in the plane. For hyperbolic functions, one can easily derive identities similar to the ones for trigonometric functions. Among many of them, we can easily see from the definition (by substituting into (5.12) and (5.13)) that

$$\cosh x = \cos(ix), \qquad i\sinh x = \sin(ix).$$

Let us notice that this snowflake is an example of an infinitely long curve which encloses a finite area. □

5.105. Calculate the series

(a) $\sum\limits_{n=1}^{\infty}\left(\frac{1}{\sqrt{n}}-\frac{1}{\sqrt{n+1}}\right)$;

(b) $\sum\limits_{n=0}^{\infty}\frac{5}{3^n}$;

(c) $\sum\limits_{n=1}^{\infty}\left(\frac{3}{4^{2n-1}}+\frac{2}{4^{2n}}\right)$;

(d) $\sum\limits_{n=1}^{\infty}\frac{n}{3^n}$;

(e) $\sum\limits_{n=0}^{\infty}\frac{1}{(3n+1)(3n+4)}$.

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

$$s_n=\frac{1}{3}+\frac{2}{3^2}+\frac{3}{3^3}+\cdots+\frac{n}{3^n},\qquad n\in\mathbb{N},$$

we immediately get that

$$\frac{s_n}{3}=\frac{1}{3^2}+\frac{2}{3^3}+\cdots+\frac{n-1}{3^n}+\frac{n}{3^{n+1}},\qquad n\in\mathbb{N}.$$

Therefore,

$$s_n-\frac{s_n}{3}=\frac{1}{3}+\frac{1}{3^2}+\frac{1}{3^3}+\cdots+\frac{1}{3^n}-\frac{n}{3^{n+1}},\qquad n\in\mathbb{N}.$$

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)

5.51. Notes. **(1)** If a power series $S(x)$ is expressed with the value of 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 our previous reasonings are still valid, we must only 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, it holds that 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 we have power series with a general center at our disposal, we can calculate the coefficients of the power series for inverse functions straightaway. We will not introduce a list of formulae here, it can easily be obtained in Maple, for instance, by the procedure "series". For illustration, we will have a look at two examples:

We have seen that

$$\mathrm{e}^x=1+x+\frac{1}{2}x^2+\frac{1}{6}x^3+\frac{1}{24}x^4+\dots.$$

Since $\mathrm{e}^0=1$, we will search for a power series centered at $x=1$ for the inverse function $\ln x$, i. e.

$$\ln x=a_0+a_1(x-1)+a_2(x-1)^2+a_3(x-1)^3+a_4(x-1)^4+\dots.$$

Applying the equality $x=\mathrm{e}^{\ln x}$, regrouping the coefficients by the powers of x and substituting, we get:

$$\begin{aligned}x&=a_0+a_1\left(x+\frac{1}{2}x^2+\frac{1}{6}x^3+\frac{1}{24}x^4+\dots\right)\\&\quad+a_2\left(x+\frac{1}{2}x^2+\dots\right)^2+a_3\left(x+\frac{1}{2}x^2+\dots\right)^3+\dots\\&=a_0+a_1x+\left(\frac{1}{2}a_1+a_2\right)x^2+\left(\frac{1}{6}a_1+a_2+a_3\right)x^3\\&\quad+\left(\frac{1}{24}a_1+\left(\frac{1}{4}+\frac{2}{6}\right)a_2+\frac{3}{2}a_3+a_4\right)x^4+\dots.\end{aligned}$$

Confronting the coefficients at the corresponding powers on both sides, we get

$$a_0=0,\ a_1=1,\ a_2=-\frac{1}{2},\ a_3=\frac{1}{3},\ a_4=-\frac{1}{4},\dots$$

which indeed corresponds to the valid expression (will be verified later):

$$\ln x=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}(x-1)^n.$$

Similarly, we can play with the series

$$\sin t=t-\frac{1}{3!}t^3+\frac{1}{5!}t^5-\frac{1}{7!}t^7+\dots$$

$$\frac{1}{(3n+1)(3n+4)} = \frac{1}{3}\cdot\frac{1}{3n+1} - \frac{1}{3}\cdot\frac{1}{3n+4}, \qquad n \in \mathbb{N}\cup\{0\},$$

which gives

$$\sum_{n=0}^{\infty}\frac{1}{(3n+1)(3n+4)} =$$

$$\lim_{n\to\infty}\frac{1}{3}\left(1-\frac{1}{4}+\frac{1}{4}-\frac{1}{7}+\frac{1}{7}-\frac{1}{10}+\cdots+\frac{1}{3n+1}-\frac{1}{3n+4}\right)$$

$$= \lim_{n\to\infty}\frac{1}{3}\left(1-\frac{1}{3n+4}\right) = \frac{1}{3}.$$

□

5.106. Verify that

$$\sum_{n=1}^{\infty}\frac{1}{n^2} < \sum_{n=0}^{\infty}\frac{1}{2^n}.$$

Solution. We can immediately see that

$$1 \le 1, \quad \frac{1}{2^2}+\frac{1}{3^2} < 2\cdot\frac{1}{2^2} = \frac{1}{2}, \quad \frac{1}{4^2}+\frac{1}{5^2}+\frac{1}{6^2}+\frac{1}{7^2} < 4\cdot\frac{1}{4^2} = \frac{1}{4},$$

or, in general:

$$\frac{1}{(2^n)^2}+\cdots+\frac{1}{(2^{n+1}-1)^2} < 2^n\cdot\frac{1}{(2^n)^2} = \frac{1}{2^n}, \qquad n\in\mathbb{N}.$$

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

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6} < 2 = \sum_{n=0}^{\infty}\frac{1}{2^n}.$$

□

5.107. 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 $+\infty$. □

5.108. Prove that the series

$$\sum_{n=0}^{\infty}\operatorname{arctg}\frac{n^2+2n+3\sqrt{n}+4}{n+1}; \qquad \sum_{n=1}^{\infty}\frac{3^n+1}{n^3+n^2-n}$$

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 $\lim\limits_{n\to\infty} a_n = 0$ for the series $\sum_{n=n_0}^{\infty} a_n$ to converge does not hold in either case. □

and the (unknown so far) series for its inverse (note that we are looking for a series centered at zero again because we have $\sin 0 = 0$)

$$\arcsin t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + \dots.$$

Substitution gives

$$t = a_0 + a_1\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 + \dots\right) +$$

$$a_2\left(t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 + \dots\right)^2 + \dots$$

$$= a_0 + a_1 t + a_2 t^2 + \left(-\frac{1}{6}a_1 + a_3\right)t3 +$$

$$\left(-\frac{2}{6}a_2 + a_4\right)t^4 + \left(\frac{1}{120}a_1 - \frac{3}{6}a_3 + a_5\right)t^5 + \dots,$$

hence

$$\arcsin t = t + \frac{1}{6}t^3 + \frac{3}{40}t^5 + \dots.$$

(3) We can also notice that if we believed right from the beginning 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 we would easily obtained the differential equation for the coefficients a_n as we know that $(x^{n+1})' = (n+1)x^n$. Therefore, from the condition that the exponential function has its derivative equal to its value at every point, it follows that

$$a_{n+1} = \frac{1}{n+1}a_n, \qquad a_0 = 1$$

and hence it is clear that $a_n = \frac{1}{n!}$.

5.109. 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 $+\infty$. □

5.110. Determine whether the series

(a) $\sum\limits_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n}$;

(b) $\sum\limits_{n=1}^{\infty} \frac{n^2+1}{n^3}$;

(c) $\sum\limits_{n=1}^{\infty} \frac{1}{n-\ln n}$

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) $\sum\limits_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n} \le \sum\limits_{n=0}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{1-\frac{1}{3}} < +\infty$;

(b) $\sum\limits_{n=1}^{\infty} \frac{n^2+1}{n^3} \ge \sum\limits_{n=1}^{\infty} \frac{n^2}{n^3} = \sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty$;

(c) $\sum\limits_{n=1}^{\infty} \frac{1}{n-\ln n} \ge \sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty$.

Hence it follows that the series (a) converges; (b) diverges to $+\infty$; (c) diverges to $+\infty$. □

More interesting exercises concerning series can be found at page 332.

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.

5.111. 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 $x \in (-\frac{1}{2}, \frac{1}{2})$ (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)

$$r = \limsup_{n\to\infty} \left|\sqrt[n]{\frac{1}{(1+i)^n}}\right| = \limsup_{n\to\infty} \left|\frac{1}{1+i}\right| = \frac{\sqrt{2}}{2},$$

□

5.112. 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) $\lim\limits_{n\to\infty} \sqrt[n]{|\,a_n\,|} = \lim\limits_{n\to\infty} \left(1+\frac{1}{n}\right)^n = \mathrm{e};$

(d) $\limsup\limits_{n\to\infty} \sqrt[n]{|\,a_n\,|} = \limsup\limits_{n\to\infty} \frac{\sqrt[n]{n^5}}{2+(-1)^n} = \limsup\limits_{n\to\infty} \frac{\left(\sqrt[n]{n}\,\right)^5}{2+(-1)^n} = 1.$

Therefore, the radius of convergence is (a) $r = 8$, (b) $r = 0$, (c) $r = 1/\mathrm{e}$, (d) $r = 1$. □

5.113. Calculate the radius r of convergence of the power series

$$\sum_{n=1}^{\infty} \mathrm{e}^{in}\, \frac{\sqrt[3]{n^3+n}\cdot 3^n}{\sqrt[3]{n^4+2n^3+1}\cdot\pi^n}\,(x-2)^n.$$

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 $r = \pi/3$. □

5.114. 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

$$\frac{1}{x-4} = -\frac{\frac{1}{4}}{1-\frac{x}{4}} = -\frac{1}{4}\left(1 + \frac{x}{4} + \frac{x^2}{4^2} + \cdots + \frac{x^n}{4^n} + \cdots\right),$$

$$\frac{1}{x+3} = \frac{\frac{1}{3}}{1-(-\frac{x}{3})} = \frac{1}{3}\left(1 - \frac{x}{3} + \frac{x^2}{3^2} + \cdots + \frac{(-x)^n}{3^n} + \cdots\right),$$

we get

$$\frac{1}{x^2-x-12} = -\frac{1}{28}\sum_{n=0}^{\infty} \frac{x^n}{4^n} - \frac{1}{21}\sum_{n=0}^{\infty} \frac{(-x)^n}{3^n} = \sum_{n=0}^{\infty} \left(\frac{(-1)^{n+1}}{21\cdot 3^n} - \frac{1}{28\cdot 4^n}\right) x^n\,.$$

□

5.115. Approximate the number sin 1° with error less than 10^{-10}.

Solution. We know that

$$\sin x = x - \frac{1}{3!}\,x^3 + \frac{1}{5!}\,x^5 - \frac{1}{7!}\,x^7 + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\,x^{2n+1}, \qquad x \in \mathbb{R}.$$

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

$$\sin 1^\circ \approx \frac{\pi}{180} - \frac{\pi^3}{180^3\cdot 3!}$$

is thus less than

$$\frac{\pi^5}{180^5\cdot 5!} < 10^{-10}.$$

□

5.116. Determine the radius r of convergence of the power series

$$\sum_{n=0}^{\infty} \frac{2^{2n}\cdot n!}{(2n)!}\, x^n\,.$$

○

5.117. Calculate the radius of convergence for $\sum_{n=1}^{\infty} 2^{\sqrt{n}}\, x^n\,.$ ○

5.118. Without calculation determine the radius of convergence of the power series

$$\sum_{n=1}^{\infty} \frac{5}{n\cdot 3^{n-1}}\, x^{n-1}\,.$$

○

5.119. Find the domain of convergence of the power series

$$\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{3^{\sqrt{n}}}\, x^n\,.$$

○

5.120. Determine for which $x \in \mathbb{R}$ the power series

$$\sum_{n=1}^{\infty} \frac{(-3)^n}{\sqrt{n^4+2n^3+111}}\, (x-2)^n$$

converges. ○

5.121. 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? ○

5.122. 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.

○

5.123. Approximate $\cos\frac{\pi}{10}$ with error less than 10^{-5}. ○

5.124. For the convergent series

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{\sqrt{n}+100},$$

bound the error of its approximation by the partial sum $s_{9\,999}$. ○

5.125. Express the function $y = \mathrm{e}^x$, 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 $a_n x^n$. ○

5.126. Find the function f to which, for $x \in \mathbb{R}$, the sequence of functions

$$f_n(x) = \frac{n^2x^3}{n^2x^2+1}, \quad n \in \mathbb{N}.$$

converges. Is this convergence uniform on $\mathbb{R}$? ○

5.127. 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? ○

5.128. 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}.

○

5.129. Determine the power series centered at $x_0 = 0$ of the function

$$f(x) = \int\limits_0^x \mathrm{e}^{t^2}\, dt, \quad x \in \mathbb{R}.$$

○

5.130. Using the integral test, find the values $a > 0$ for which the series

$$\sum_{n=1}^{\infty} \frac{1}{n^a}$$

converges. ○

5.131. 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. ○

5.132. Determine all $x \in \mathbb{R}$ for which the power series $\sum\limits_{i=1}^{\infty} \frac{x^{2n}}{n^2}$ is convergent. ○

Solution. For $x \in [-1, 1]$. □

5.133. For which $x \in \mathbb{R}$ does the series

$$\sum_{n=1}^{\infty} \frac{\ln(n!)}{n^x}$$

converge? ○

5.134. Determine whether the series

$$\sum_{n=1}^{\infty} (-1)^{n-1} \tan \tfrac{1}{n\sqrt{n}}$$

converges absolutely, converges conditionally, diverges to $+\infty$, diverges to $-\infty$, or none of the above. (such a series is sometimes said to be oscillating). ○

5.135. Calculate the series

$$\sum_{n=1}^{\infty} \tfrac{1}{n \cdot 3^n}$$

with the help of an appropriate power series. ○

5.136. For $x \in (-1, 1)$, add

$$x - 4x^2 + 9x^3 - 16x^4 + \cdots$$

○

5.137. Supposing $|\,x\,| < 1$, determine the series

(a) $\sum\limits_{n=1}^{\infty} \tfrac{1}{2n-1}\, x^{2n-1}$;

(b) $\sum\limits_{n=1}^{\infty} n^2 x^{n-1}$.

○

5.138. Calculate

$$\sum_{n=1}^{\infty} \tfrac{2n-1}{(-2)^{n-1}}$$

using the power series

$$\sum_{n=0}^{\infty} (-1)^n \,(2n+1)\, x^{2n}$$

for some $x \in (-1, 1)$. ○

5.139. For $x \in \mathbb{R}$, calculate the series

$$\sum_{n=0}^{\infty} \tfrac{1}{2^n \cdot n!}\, x^{3n+1}.$$

○

J. Additions into the ZOO

5.140. Determine the maximal subset of $\mathbb{R}$ where the function

$$y = \operatorname{arctg}\,\left(x^{21} + \sin x\right) \cdot \tfrac{\mathrm{e}^{\cos\left(\sqrt[5]{x}-21+\cos\, x\right)+x-256x^3} - 11}{2+x^{252}}$$

can be defined. ○

Solution. $\mathbb{R}$.

5.141. Write the maximal domain of the function

$$y = \tfrac{\arccos\,(\ln x)}{\sqrt{x^2-1}}.$$

○

Solution. $(1, \mathrm{e}]$.

5.142. Determine the domain and the range of the function

$$y = \tfrac{x-1}{2-3x}.$$

Then determine the function inverse to this one. ○

Solution. $\left(-\infty, \frac{2}{3}\right) \cup \left(\frac{2}{3}, +\infty\right)$; $\left(-\infty, -\frac{1}{3}\right) \cup \left(-\frac{1}{3}, +\infty\right)$; $y = \frac{2x+1}{3x+1}$, $x \neq -\frac{1}{3}$.

5.143. 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) $y = \sinh x = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{2}$;
(h) $y = \cosh x = \frac{\mathrm{e}^x + \mathrm{e}^{-x}}{2}$

with the maximal domain odd? ○

Solution. (a) yes; (b) no; (c) no; (d) no; (e) yes; (f) yes; (g) yes; (h) no.

5.144. 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) $y = \sinh x = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{2}$;
(h) $y = \cosh x = \frac{\mathrm{e}^x + \mathrm{e}^{-x}}{2}$

with the maximal domain even? ○

5.145. 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 and whether it is even. ○

5.146. Is the function

(a) $y = \ln(\cos x)$;
(b) $y = \tan\,(3x) + 2\sin(6x)$

with maximal domain periodic? ○

5.147. Draw the graphs of the functions

$$f(x) = \mathrm{e}^{|\,x\,|}, \quad x \in \mathbb{R}; \qquad g(x) = \ln|\,x\,|, \quad x \in \mathbb{R} \setminus \{0\}.$$

○

5.148. Draw the graph of the function

$$y = 2^{-|\,x\,|}, \quad x \in \mathbb{R}.$$

○

5.149. The functions

$$\sinh x = \tfrac{\mathrm{e}^x - \mathrm{e}^{-x}}{2}, \quad x \in \mathbb{R}; \qquad \cosh x = \tfrac{\mathrm{e}^x + \mathrm{e}^{-x}}{2}, \quad x \in \mathbb{R};$$

$$\tanh x = \tfrac{\sinh x}{\cosh x}, \quad x \in \mathbb{R}; \qquad \coth x = \tfrac{\cosh x}{\sinh x}, \quad x \in \mathbb{R} \setminus \{0\}$$

are called hyperbolic functions. Determine the derivatives of these functions on their domains. ○

5.150. 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 $\mathbb{R}$. ○

Note: 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

$$(\operatorname{arcosh} x)' = \tfrac{1}{\sqrt{x^2-1}}, \quad x > 1,$$

$$(\operatorname{artanh} x)' = \tfrac{1}{1-x^2}, \quad |\,x\,| < 1,$$

$$(\operatorname{arcoth} x)' = \tfrac{1}{1-x^2}, \quad |\,x\,| > 1.$$

5.151. Calculate:

$$2 + 1 + \frac{2}{2!} + \frac{1}{3!} + \frac{2}{4!} + \frac{1}{5!} + \frac{2}{6!} + \cdots$$

○

Solution. Confronting the series with the expansions of the functions sinh and cosh into power series, we get the result

$$\sinh(1) + 2\cosh(1).$$

□

K. Additional exercises to the whole chapter

5.152. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(1) = 1$, $P(2) = 28$, $P(0) = 2$, $P'(0) = 1$, $P'(1) = 9$. ○

5.153. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(0) = 0$, $P(1) = 4$, $P(-1) = -2$, $P'(0) = 1$, $P'(1) = 7$. ○

5.154. Determine a polynomial $P(x)$ of the least degree possible satisfying the conditions $P(0) = -1$, $P(1) = -1$, $P'(-1) = 10$, $P'(0) = -1$, $P'(1) = 6$. ○

5.155. From the definition of a limit, prove that

$$\lim_{x\to 0}\left(x^3 - 2\right) = -2.$$

○

5.156. 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. ○

5.157. From the definition of a limit, show that

$$\lim_{x\to -\infty}\frac{3\,(x-2)^4}{2} = +\infty.$$

○

5.158. Determine both one-sided limits

$$\lim_{x\to 0+}\arctan\frac{1}{x}, \qquad \lim_{x\to 0-}\arctan\frac{1}{x}.$$

Knowing the result, decide existence of the limit

$$\lim_{x\to 0}\arctan\frac{1}{x}.$$

○

5.159. Do the following limits exist?

$$\lim_{x\to 0}\frac{\sin x}{x^3}, \qquad \lim_{x\to 0}\frac{5x^4+1}{x}$$

○

5.160. Calculate the limit

$$\lim_{x\to 0}\frac{\tan x - \sin x}{\sin^3 x}.$$

○

5.161. Determine

$$\lim_{x\to \pi/6}\frac{2\sin^3 x + 7\sin^2 x + 2\sin x - 3}{2\sin^3 x + 3\sin^2 x - 8\sin x + 3}.$$

○

5.162. For any $m, n \in \mathbb{N}$, determine

$$\lim_{x\to 1}\frac{x^m - 1}{x^n - 1}.$$

○

5.163. Calculate

$$\lim_{x\to+\infty}\left(\sqrt{x^2+x}-x\right).$$

○

5.164. Determine

$$\lim_{x\to+\infty}\left(x\sqrt{1+x^2}-x^2\right).$$

○

5.165. Calculate

$$\lim_{x\to 0}\frac{\sqrt{2}-\sqrt{1+\cos x}}{\sin^2 x}.$$

○

5.166. Determine

$$\lim_{x\to 0}\frac{\sin(4x)}{\sqrt{x+1}-1}.$$

○

5.167. Calculate

$$\lim_{x\to 0-}\frac{\sqrt{1+\tan x}-\sqrt{1-\tan x}}{\sin x}.$$

○

5.168. Calculate

$$\lim_{x\to-\infty}\frac{2^x+\sqrt{1+x^2-x^9}-7x^5+44x^2}{3^x+\sqrt[5]{6x^6+x^2}-18x^5-592x^4}.$$

○

5.169. 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 $g:\mathbb{R}\to\mathbb{R}$? ○

5.170. Determine the limit

$$\lim_{n\to\infty}\left(\frac{n}{n+5}\right)^{2n-1}.$$

○

5.171. Calculate

$$\lim_{x\to 0-}\frac{\sin x-x}{x^3}.$$

○

5.172. For $x>\mathrm{e}$, determine the sign of the derivative of the function

$$f(x)=\arctan\tfrac{\ln x}{-1+\ln x}.$$

○

Solution. $f'(x)<0,\ x>\mathrm{e}$.

5.173. Determine all local extrema of the function

$$y=x\ln^2 x$$

defined on the interval $(0, +\infty)$. ○

Solution. The function has a local maximum at the point $x_1 = \mathrm{e}^{-2}$ and it has a local minimum at the point $x_2 = 1$.

5.174. Is there a real number a such that the function $y = ax + \sin x$ has a global minimum on the interval $[0, 2\pi]$ at the point $x_0 = 5\pi/4$? ○

Solution. There is not: for $a = \sqrt{2}/2$, there is only a local extremum at the point.

5.175. Find the absolute minimum of the function

$$y = \mathrm{e}\,x - \ln x, \quad x > 0$$

on its domain. ○

Solution. $2 = \mathrm{e}\,\frac{1}{\mathrm{e}} - \ln\frac{1}{\mathrm{e}}$.

5.176. Determine the maximum value of the function

$$y = \sqrt[3]{3x}\,\mathrm{e}^{-x}, \quad x \in \mathbb{R}.$$

○

Solution. $\frac{1}{\sqrt[3]{\mathrm{e}}}$.

5.177. Find the absolute extrema of the polynomial $p(x) = x^3 - 3x + 2$ on the interval $[-3, 2]$. ○

Solution. $4 = p(-1) = p(2)$, $-16 = p(-3)$.

5.178. Let a moving object's distance 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 distance is in meters. Determine

(a) the initial (i. e. at the time $t = 0$ s) speed of the object;
(b) the time and position at which its speed is zero;
(c) its speed and acceleration at the time $t = 4$ s.

Let us remark that the object's speed is the derivative of its position and acceleration is the derivative of its speed. ○

5.179. 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 $x_0 > 0$. ○

5.180. Determine whether the derivative of the function

$$f(x) = x \arctan\tfrac{1}{x}, \quad x \in \mathbb{R} \setminus \{0\}, \qquad f(0) = 0$$

at the point $x_0 = 0$ exists. ○

5.181. Does the derivative of the function

$$y = \sin\left(\arctan\left(\left|\,12x^{21} + 11\,\right| \cdot \tfrac{\mathrm{e}^{\cos(x+2)-x^3}}{-11-x^{12}}\right)\right) + \sin(\sin(\sin(\sin x))), \quad x \in \mathbb{R}$$

at the point $x_0 = \pi^3 + 3^\pi$ exist? ○

5.182. Determine whether the derivative of the function

$$f(x) = \left(x^2 - 1\right)\sin\tfrac{1}{x+1}, \quad x \neq -1\ (x \in \mathbb{R}), \qquad f(-1) = 0$$

at the point $x_0 = -1$ exists. ○

5.183. 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 $x_1 = 5$, $x_2 = 9$. ○

5.184. Find functions f and g which have derivatives at no real point, yet their composition $f \circ g$ is differentiable at every point of the real line. ○

5.185. Using the 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) $y = \frac{1+x-x^2}{1-x+x^2}, \quad x \in \mathbb{R}.$

○

5.186. By any means, determine the derivative of the function

(a) $y = \sqrt{x\sqrt{x\sqrt{x}}}, \quad x \in (0, +\infty);$
(b) $y = \ln\left|\tan\frac{x}{2}\right|, \quad x \in \mathbb{R} \setminus \{n\pi;\ n \in \mathbb{Z}\}.$

○

5.187. Write the derivative of the function

$$y = \sin\left(\sin\left(\sin x\right)\right), \quad x \in \mathbb{R}.$$

○

5.188. For the function

$$f(x) = \arccos\frac{1-x}{\sqrt{2}} + \sqrt[3]{x^3}$$

having the maximum possible domain, calculate f' on the largest subset of $\mathbb{R}$ where this derivative exists. ○

5.189. At any point $x \notin \{n\pi;\ n \in \mathbb{Z}\}$, determine the first derivative of the function $y = \sqrt[3]{\sin x}$. ○

5.190. For $x \in \mathbb{R}$, differentiate

$$x\sqrt{1+x^2} + \mathrm{e}^x\left(x^2 - 2x + 2\right).$$

○

5.191. Calculate $f'(1)$ if

$$f(x) = (x-1)(x-2)^2(x-3)^3, \quad x \in \mathbb{R}.$$

○

5.192. Determine the derivative of the function

$$y = \sqrt[3]{\frac{1+x^3}{1-x^3}}, \quad |\,x\,| \neq 1 \ (x \in \mathbb{R}).$$

○

5.193. 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. ○

5.194. Determine f' on a maximal set if $f(x)=\log_x \mathrm{e}$. ○

5.195. Express the derivative of the product of four functions

$$\left[f(x)g(x)h(x)k(x)\right]'$$

as a sum of products of their derivatives and themselves, supposing all of these functions are differentiable. ○

5.196. Determine the derivative of the function

$$y=\frac{x^3\,(x+1)^2\sqrt[3]{x+2}}{(x+3)^2}$$

for $x>0$. ○

5.197. 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 and which is approaching it 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 $x'(t)\le 0$, $y'(t)\ge 0$ 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 $x'(t)\le 0$. Similarly we can get that $y'(t)\ge 0$ and also $s'(t)\le 0$. 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

$$s\,(t_0)=5,\qquad s'\,(t_0)=-160,\qquad y'\,(t_0)=120$$

and that ($s(t)$ is the hypotenuse of the right triangle)

$$(x(t)-y(t))^{\;2}+3^2=s^2\,(t). \tag{5.9}$$

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 (‖5.9‖), we get

$$2\,(x(t)-y(t))\,\left(x'\,(t)-y'\,(t)\right)=2s(t)s'\,(t)$$

and then for $t=t_0$,

$$2\cdot 4\left(x'\,(t_0)-120\right)=2\cdot 5\cdot(-160),\quad \text{i. e.}\quad x'\,(t_0)=-80.$$

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. □

5.198. 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

$$P'(x) = 3ax^2 + 2ax + 1$$

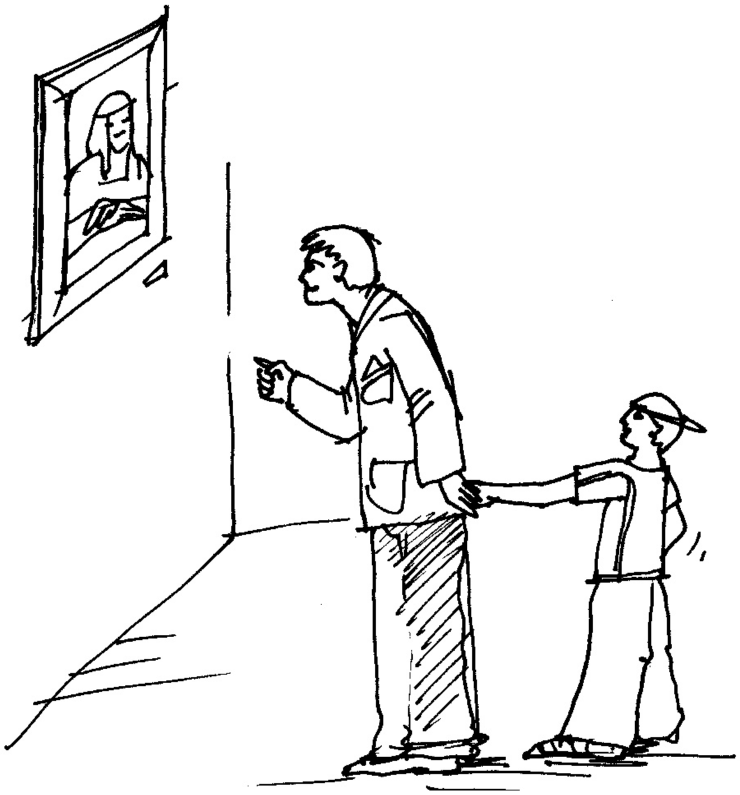
and it will keep the sign iff the discriminant is non-positive. Thus we get the condition

$$\begin{aligned} 4a^2 - 12a &\leq 0 \\ 4a(a-3) &\leq 0, \end{aligned}$$

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 $(0, 3]$. □

5.199. 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 $h \in (a, b)$).

>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

$$\tan\varphi = \tan\,(\alpha - \beta) = \tfrac{\tan\alpha - \tan\beta}{1+\tan\alpha\,\tan\beta} = \tfrac{\frac{b-h}{x} - \frac{a-h}{x}}{1+\frac{b-h}{x}\cdot\frac{a-h}{x}} = \tfrac{x(b-a)}{x^2+(b-h)(a-h)}.$$

So it suffices to find the global maximum of the function

$$f(x) = \tfrac{x(b-a)}{x^2+(b-h)(a-h)}, \qquad x \in [0, +\infty).$$

From the expression

$$f'(x) = \tfrac{(b-a)\left[x^2+(b-h)(a-h)\right]-2x^2(b-a)}{\left[x^2+(b-h)(a-h)\right]^2} = \tfrac{(b-a)\left[(b-h)(a-h)-x^2\right]}{\left[x^2+(b-h)(a-h)\right]^2}, \quad x \in (0, +\infty),$$

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)

$$\frac{y_1+y_2}{2} \geq \sqrt{y_1\, y_2}, \qquad y_1, y_2 \geq 0,$$

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 $\approx 11.5\,^\circ$ 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 $\approx 45.6\,^\circ$ whereas his son has $\varphi_0 \doteq 0.339\,8$ rad $\approx 19.5\,^\circ$. 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. □

5.200. 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 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

$$\text{(5.10)} \qquad \frac{\sin\varphi_1}{\sin\varphi_2} = \frac{v_1}{v_2}.$$

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 (‖5.10‖), 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 g m^{-3}) is $n \doteq 1.000272$. 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 (‖5.10‖) might not exist. This corresponds to the total reflection (there is no refracted light at all). □

5.201. The rainbow. Why is the rainbow circular?

Solution. In the exercise called Snell's law we clarified what is the rainbow caused by. (It is created by sunlight being refracted while entering a droplet of water.) Now we will go on with this problem. To be concrete, we will examine how the rays behave when going through the droplets. (See the picture.) 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 $\mid OA \mid = \mid OB \mid$, the angle of reflection equals φ_r. Here as well, of course, some 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. Let us add that we omit the case of the so-called secondary rainbow arc, i. e. when the ray reflects twice inside the droplet before refracting out of it.

We will express the angle $\alpha := \measuredangle AIC$. Since $\measuredangle OAI = \varphi_i$ and $\measuredangle OAB = \varphi_r$, we get $\measuredangle BAI = \varphi_i - \varphi_r$. Then

$$\measuredangle BIA = \pi - (\measuredangle ABI) - (\measuredangle BAI) = \pi - (\pi - \varphi_r) - (\varphi_i - \varphi_r) = 2\varphi_r - \varphi_i$$

and further

$$\alpha = 2 \cdot \measuredangle BIA = 4\varphi_r - 2\varphi_i .$$

By Snell's law, we have

$$\tfrac{\sin\varphi_i}{\sin\varphi_r} = n,$$

where n stands for the refractive index for water (as we assume that air's index of refraction equals 1). Thus we have that

$$\varphi_r = \arcsin \tfrac{\sin\varphi_i}{n}$$

whence it follows that

$$\alpha = 4\arcsin\left(\frac{\sin\varphi_i}{n}\right) - 2\varphi_i . \tag{5.11}$$

For the rays going out of the droplet, the value 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 center of the droplet, then

$$\sin\varphi_i = \frac{y}{R} \quad \text{for} \quad y \in [0, R]. \tag{5.12}$$

Of course, we can assume (thanks to the huge distance of the Sun from Earth) 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 considered values y for sufficiently small $\delta > 0$. It thus makes sense to analyze the function (see (‖5.11‖) and (‖5.12‖))

$$\alpha(y) = 4\arcsin \tfrac{y}{nR} - 2\arcsin \tfrac{y}{R}, \qquad y \in [0, R].$$

By selecting the appropriate unit of length (for which $R = 1$) we can turn to the function

$$\alpha(x) = 4\arcsin \tfrac{x}{n} - 2\arcsin x, \qquad x \in [0, 1].$$

Having calculated the derivative

$$\alpha'(x) = \tfrac{4}{n\sqrt{1-\frac{x^2}{n^2}}} - \tfrac{2}{\sqrt{1-x^2}}, \qquad x \in (0, 1),$$

we can easily determine that the equation $\alpha'(x) = 0$ has a unique solution

$$x_0 = \sqrt{\tfrac{4-n^2}{3}} \in (0,1), \quad \text{if} \quad n^2 \in (1,4).$$

Let us set $n = 4/3$ (which is approximately the refractive index of water). Further,

$$\alpha'(x) > 0, \quad x \in (0, x_0), \qquad \alpha'(x) < 0, \quad x \in (x_0, 1).$$

We have found that at the point

$$x_0 = \sqrt{\frac{4-\left(\frac{4}{3}\right)^2}{3}} = \tfrac{2}{3}\sqrt{\tfrac{5}{3}} \doteq 0.86,$$

the function α has a global maximum

$$\alpha(x_0) = 4 \arcsin \tfrac{\sqrt{5}}{2\sqrt{3}} - 2 \arcsin \tfrac{2\sqrt{5}}{3\sqrt{3}} \doteq 0.734 \text{ rad} \approx 42\,^\circ.$$

Although it is amazing that the peak of the rainbow cannot be above the level of approximately $42\,^\circ$ with regard to the observer, what is even more amazing are the values

$$\alpha(0.74) \doteq 39.4\,^\circ, \quad \alpha(0.94) \doteq 39.2\,^\circ, \quad \alpha(0.8) \doteq 41.2\,^\circ, \quad \alpha(0.9) \doteq 41.5\,^\circ.$$

Those imply (the function 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 $39\,^\circ$ to around $42\,^\circ$, and 10 % lie in a band thinner than $1\,^\circ$. Furthermore, if we consider

$$\alpha(0.84) \doteq 41.9\,^\circ, \qquad \alpha(0.88) \doteq 41.9\,^\circ,$$

we can see that the rays for which α is close to $42\,^\circ$ have the greatest intensity. Let us emphasize that this is an instance of the so-called principle of minimum deviation: the highest concentration of the diffused light happens to be 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 the 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 picture). Thus when the sun is setting, the rainbow has the shape of a semicircle. Let us remark that the rainbow exists only with regard to its observer – it is not anchored in the space. Eventually, let us add that the circular shape of the rainbow was examined as early as 1635–1637 by René Descartes. □

5.202. L'Hospital's pulley.

A rope of length r is tied to the ceiling at point A. A pulley is attached to its other end. Another rope of length $l > \sqrt{d^2 + r^2}$, going through the pulley, is tied to the ceiling at point B which is at distance d from the point A. A weight is attached to this rope. In what position will the weight stabilize (the system will be in a stationary position)? Omit the mass and the size of the ropes and the pulley. See the picture.

Solution. The system will be in a stationary position if its potential energy is minimized, i. e. the distance $f(x)$ of the weight from the ceiling is maximal. However, this means that for $r \geq d$, the pulley only moves under the point B. Further on we will thus suppose that $r < d$. By the Pythagorean theorem, the distance of the pulley from the ceiling is $\sqrt{r^2 - x^2}$ and from the weight then $l - \sqrt{(d-x)^2 + r^2 - x^2}$, which gives

$$f(x) = \sqrt{r^2 - x^2} + l - \sqrt{(d-x)^2 + r^2 - x^2}.$$

The state of the system is fully given by the value $x \in [0, r]$ (see the picture), so it suffices to find the global maximum of the function f on the interval $[0, r]$.

First, we calculate the derivative

$$f'(x) = \frac{-x}{\sqrt{r^2-x^2}} - \frac{-(d-x)-x}{\sqrt{(d-x)^2+r^2-x^2}} = \frac{-x}{\sqrt{r^2-x^2}} + \frac{d}{\sqrt{(d-x)^2+r^2-x^2}}, \qquad x \in (0,r).$$

Exponentiating the equatino $f'(x) = 0$ for $x \in (0,r)$ leads to

$$\frac{x^2}{r^2-x^2} = \frac{d^2}{(d-x)^2+r^2-x^2}.$$

Multiplying both sides by $\left(r^2-x^2\right)\left((d-x)^2+r^2-x^2\right)$ then leads to

$$2dx^3 - \left(2d^2+r^2\right)x^2 + d^2r^2 = 0, \qquad x \in (0,r).$$

If we notice that one of the roots of the left-hand polynomial is $x = d$, we can easily transform the last equation into the form

$$(x-d)\left(2dx^2 - r^2x - dr^2\right) = 0, \qquad x \in (0,r),$$

or (we have a formula for the quadratic equation)

$$2d\,(x-d)\left(x - \frac{r^2+r\sqrt{r^2+8d^2}}{4d}\right)\left(x - \frac{r^2-r\sqrt{r^2+8d^2}}{4d}\right) = 0, \qquad x \in (0,r).$$

Hence we can see that the equation $f'(x) = 0$ has at most one solution on the interval $(0,r)$. (Since $r < d$ and $\sqrt{r^2+8d^2} > r$, there are surely not two roots of the considered polynomial in x in the interval $(0,r)$.) It remains to determine whether

$$x_0 = \frac{r^2+r\sqrt{r^2+8d^2}}{4d} = \frac{1}{4}r\left[\frac{r}{d} + \sqrt{\left(\frac{r}{d}\right)^2+8}\right] \in (0,r).$$

Realizing that $r, d > 0$ and $r < d$, we get

$$0 < x_0 < \frac{1}{4}r\left[1+\sqrt{1^2+8}\right] = r.$$

As the function f' is continuous on the interval $(0,r)$, it can change sign only at the point 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 the global maximum on the interval $[0,r]$ at the point x_0. □

5.203. A nameless mail company can only transport parcels whose length does not exceed 108 inches and whose sum of length and maximal perimeter is at most 165 inches. Find the largest (i. e. having the greatest volume) parcel which can be transported by this company.

Solution. Let M denote the value 165 (inches) and x the parcel's length (in inches as well). Apparently, the wanted parcel has such a shape that for any $t \in (0,x)$, its cross section has a constant perimeter (the maximal one). We will denote this perimeter by p (in inches, again). We want the parcel to have the greatest volume so that the cross section of a given perimeter has the greatest area possible. It is not difficult to realize that the largest planar figure of a given perimeter is a disc. Thus we have derived that the desired parcel has the shape of a cylinder with height equal to x and radius $r = p/2\pi$.

Its volume is

$$V = \pi r^2 x = \frac{p^2x}{4\pi},$$

and it must be that $p + x \leq M$ and $x \leq 108$. Thus we consider the parcel for which $p + x = M$. Its volume is

$$V(x) = \frac{(M-x)^2x}{4\pi} = \frac{x^3-2Mx^2+M^2x}{4\pi} \quad \text{where} \quad x \in (0,108]\,.$$

Having calculated the derivative

$$V'(x) = \tfrac{3x^2-4Mx+M^2}{4\pi} = \tfrac{3(x-M)\left(x-\frac{M}{3}\right)}{4\pi}, \qquad x \in (0, 108)\,,$$

we easily find out the 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

$$V\left(\tfrac{M}{3}\right) = \tfrac{M^3}{27\pi} \doteq 0.011\,789\,M^3 \approx 0.867\,8 \text{ m}^3.$$

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. It suffices to realize that 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 square of the ratio.) Thus the parcel's volume is the function

$$V(x) = Sx = kp^2x = k\,(M-x)^2\,x, \qquad x \in (0, 108]\,.$$

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

$$V = a^2x = \tfrac{M^3}{6^2\cdot 3} \doteq 0.009\,259\,M^3 \approx 0.681\,6 \text{ m}^3.$$

For a parcel in the shape of a ball (when x is the diameter), the condition $p + x \leq M$ can immediately be expressed as $\pi x + x \leq M$, i. e. $x \leq M/(\pi + 1) < 108$. Thus for $x = M/(\pi + 1)$, we get the maximal volume

$$V = \tfrac{4}{3}\pi\left(\tfrac{x}{2}\right)^3 = \tfrac{\pi M^3}{6(\pi+1)^3} \doteq 0.007\,370\,M^3 \approx 0.542\,6 \text{ m}^3.$$

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 $x = M/5$ we get the maximal volume

$$V = x^3 = \left(\tfrac{M}{5}\right)^3 = 0.008\,M^3 \approx 0.588\,9 \text{ m}^3.$$

Let us add that the length of the edges of the cube which has the same volume as the found cylinder is

$$a = \tfrac{M}{3\sqrt[3]{\pi}} \doteq 0.227\,595\,M \approx 0.953\,849 \text{ m}.$$

Let us realize its length and perimeter sum to $5a \doteq 1.138\,M$, i. e. more than the company's limit by around 14 %. □

5.204. A large military area (further denoted by MA) having the shape of a square and area of 100 km^2 is bounded along its perimeter by a narrow path. From the starting point in one corner of MA, one can get to the target point inside MA by going 5 km along the path and then 2 km perpendicularly to it. However, one can also go along the path at 5 kph for any time period and then askew through the MA at 3 kph. What distance do you have to travel along the path if you want to get there as soon as possible?

Solution. To travel x km along the path (where $x \in [0, 5]$), we need $x/5$ hours. Our way through MA will then be

$$\sqrt{2^2 + (5-x)^2} = \sqrt{x^2 - 10x + 29}$$

kilometers long and we will cover it in $\sqrt{x^2 - 10x + 29}/3$ hours. Altogether, our journey will take

$$f(x) = \tfrac{1}{5}x + \tfrac{1}{3}\sqrt{x^2 - 10x + 29}$$

hours (let us remind that $x \in [0, 5]$). The only zero point of the function

$$f'(x) = \tfrac{1}{5} + \tfrac{1}{3}\,\tfrac{x-5}{\sqrt{x^2-10x+29}}$$

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 we should go 3.5 km along the path. □

5.205. You find yourself 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 picture). 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 apparently given by first rowing straight 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 picture), 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 will take

$$\tfrac{\sqrt{d^2+x^2}}{v_1} \text{ hours}$$

and the final run then

$$\tfrac{l-x}{v_2} \text{ hours.}$$

We want to minimize the total time, i. e. the function

$$t(x) = \tfrac{\sqrt{d^2+x^2}}{v_1} + \tfrac{l-x}{v_2}$$

on the interval $[0, l]$. Further, we can assume that $v_1 < v_2$. (Clearly for $v_1 \geq v_2$ the optimal strategy is to row straight to the target point, which corresponds to $x = l$.)

First, we calculate the first derivative

$$t'(x) = \tfrac{x}{v_1\sqrt{d^2+x^2}} - \tfrac{1}{v_2}, \qquad x \in (0, l)$$

and then the second derivative

$$t''(x) = \tfrac{d^2}{v_1\sqrt{(d^2+x^2)^3}}, \qquad x \in (0, l).$$

Further, we solve the equation

$$t'(x) = 0, \qquad \text{i. e.} \qquad \tfrac{x}{\sqrt{d^2+x^2}} = \tfrac{v_1}{v_2}.$$

Exponentiating this equation gives

$$x^2 = \left(\tfrac{v_1}{v_2}\right)^2 \left(d^2 + x^2\right).$$

Simple rearrangements lead to

$$x^2 = \frac{\left(\frac{v_1}{v_2}\right)^2 d^2}{1-\left(\frac{v_1}{v_2}\right)^2}, \qquad \text{i. e.} \qquad x = \frac{\frac{v_1}{v_2}\,d}{\sqrt{1-\left(\frac{v_1}{v_2}\right)^2}}.$$

Let us realize that we consider only $x \in (0, l)$. Thus we are interested in whether

$$\frac{\frac{v_1}{v_2}\,d}{\sqrt{1-\left(\frac{v_1}{v_2}\right)^2}} < l, \qquad \text{i. e.}\ \frac{v_1}{v_2} < \frac{l}{\sqrt{l^2+d^2}}.$$

If this inequality holds, then also $v_1 < v_2$ and the function t' changes sign only at the point

$$x_0 = \frac{\frac{v_1}{v_2}\,d}{\sqrt{1-\left(\frac{v_1}{v_2}\right)^2}} \in (0, l),$$

and this change is from negative to positive (consider $\lim_{x\to 0+} t'(x) < 0$ and $t''(x) > 0$, $x \in (0, l)$). This means that in this case, at the point x_0 there is the global minimum of the function t on the interval $[0, l]$. However, if the inequality (‖5.205‖) is false, then we have $t'(x) < 0$ for all $x \in (0, l)$ whence it follows that the global minimum of the function t on $[0, l]$ is at the right-hand marginal point (the function t is decreasing on its domain). The fastest journey will take (in hours)

$$t\left(x_0\right) = \frac{\sqrt{d^2 + x_0^2}}{v_1} + \frac{l - x_0}{v_2} = \frac{1}{v_1}\sqrt{\text{[redacted]}}\; d^2 + \frac{\left(\frac{v_1}{v_2}\right)^2 d^2}{1 - \left(\frac{v_1}{v_2}\right)^2} + \frac{1}{v_2}\left(l - \frac{\frac{v_1}{v_2}\, d}{\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2}}\right) = \frac{d}{v_1\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2}} + \frac{l\sqrt{1 - \left(\frac{v_1}{v_2}\right.}}{v_2\sqrt{1 - \left(\right.}}$$

$$= \frac{dv_2 + lv_1\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2} - \frac{v_1^2}{v_2}\, d}{v_1 v_2\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2}} = \frac{dv_2\left(1 - \left(\frac{v_1}{v_2}\right)^2\right) + lv_1\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2}}{v_1 v_2\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2}} = \frac{dv_2\sqrt{1 - \left(\frac{v_1}{v_2}\right)^2} + lv_1}{v_1 v_2} = \frac{d\sqrt{v_2^2}}{v_1 v}$$

supposing (‖5.205‖), and

$$t\left(l\right) = \frac{\sqrt{d^2 + l^2}}{v_1}\ \text{hours}$$

if (‖5.205‖) does not hold. □

5.206. 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 to equal 2 400 m?

Solution. Let us reformulate the statement of the problem: We want to maximize the product $5ab$ while satisfying the condition

$$6b + 10a = 2\,400, \quad a, b > 0. \tag{5.13}$$

It can easily be shown that the function

$$a \mapsto 5a\,\tfrac{2\,400 - 10a}{6}$$

defined for $a \in [0, 240]$ takes the maximal value at the point $a = 120$. Hence the result is

$$a = 120 \text{ m}, \qquad b = 200 \text{ m}.$$

Let us add that the mentioned value of b immediately follows from (‖5.13‖). □

5.207. 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?

5.208. Choose 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 spare the least amount of paint possible to prime its bottom and walls. ○

5.209. Express the number 28 as a sum of two non-negative numbers such that the sum of the first summand squared and the second summand cubed is as small as possible. ○

5.210. With the help of the first derivative, find the real number $a > 0$ for which the sum $a + 1/a$ is minimal. Then solve this problem without using the differential calculus. ○

5.211. Inscribe a rectangle with the greatest perimeter possible into a semidisc with radius r. Determine the rectangle's perimeter. ○

5.212. Among the rectangles with perimeter $4c$, find the one having the greatest area (if such one exists) and determine the lengths of its sides. ○

5.213. Find the height h and the radius r of the largest (i. e. having the greatest volume) cone which fits into a ball of radius R. ○

5.214. From the triangles with a given perimeter p, select the one with the greatest area. ○

5.215. On the parabola given by the equation $2x^2 - 2y = 9$, find the points which are closest to the origin of the coordinate system. ○

5.216. Your task is to create a one-liter 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. ○

5.217. Determine the distance of the point $[3, -1] \in \mathbb{R}^2$ from the parabola $y = x^2 - x + 1$. ○

5.218. Determine the distance of the point $[-4, -2] \in \mathbb{R}^2$ from the parabola $y = x^2 + x + 1$. ○

5.219. At the 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 the point $B = [-2, -1]$ 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 their distance be at that moment? ○

5.220. 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? ○

5.221. Determine the maximum possible volume of a cone with surface area 3π cm^2 (the surface area of its base is included as well). The area of a cone is $P = \pi r(r + h)$, its volume then $V = \frac{1}{3}\pi r^2 h$, where r is the radius of its base and h is its height. ○

5.222. 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 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?

○

5.223. Suppose you own an excess of funds without the possibility to invest outside your own factory which acts at 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? ○

5.224. Determine

$$\lim_{x\to 0}\left(\cot x - \frac{1}{x}\right).$$

Solution. If we realize 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 $\infty - \infty$. We can thus consider the (two-sided) limit.

We will write the cotangent function as the ratio of the cosine and the sine and convert the fractions to a common denominator, i. e.

$$\lim_{x\to 0}\left(\cot x - \frac{1}{x}\right) = \lim_{x\to 0}\frac{x\cos x - \sin x}{x\sin x}.$$

Thus we have obtained 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 then get

$$\lim_{x\to 0}\frac{-x\sin x}{\sin x + x\cos x} = \lim_{x\to 0}\frac{-\sin x - x\cos x}{\cos x + \cos x - x\sin x} = \frac{0-0}{1+1-0} = 0.$$

□

5.225. Determine the limit

$$\lim_{x\to 1-}(1-x)\tan\frac{\pi x}{2}.$$

○

5.226. Calculate

$$\lim_{x\to \frac{\pi}{2}-}\left(\frac{\pi}{2} - x\tan x\right).$$

○

5.227. Using l'Hospital's rule, determine

$$\lim_{x\to +\infty}\left(\left(3^{\frac{1}{x}} - 2^{\frac{1}{x}}\right)x\right).$$

○

5.228. Calculate

$$\lim_{x\to 1}\left(\frac{1}{2\ln x} - \frac{1}{x^2-1}\right).$$

○

5.229. By l'Hospital's rule, calculate the limit

$$\lim_{x\to +\infty}\left(\cos\frac{2}{x}\right)^{x^2}.$$

○

5.230. Determine

$$\lim_{x\to 0} (1-\cos x)^{\sin x} = \dots$$

○

5.231. 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. ○

5.232. By any means, verify that

$$\lim_{x\to 0} \frac{\mathrm{e}^x - 1}{x} = 1.$$

○

5.233. By applying the ratio test (also called D'Alembert's criterion; see 5.46), determine whether the infinite series

(a) $\sum\limits_{n=1}^{\infty} \frac{2^n\cdot(n+1)^3}{3^n}$;

(b) $\sum\limits_{n=1}^{\infty} \frac{6^n}{n!}$;

(c) $\sum\limits_{n=1}^{\infty} \frac{n^n}{n^2\cdot n!}$

converges.

Solution. Since ($a_n \geq 0$ for all n)

(a) $\lim\limits_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim\limits_{n\to\infty} \frac{2^{n+1}\cdot(n+2)^3\cdot 3^n}{3^{n+1}\cdot 2^n\cdot(n+1)^3} = \lim\limits_{n\to\infty} \frac{2(n+2)^3}{3(n+1)^3} = \lim\limits_{n\to\infty} \frac{2n^3}{3n^3} = \frac{2}{3} < 1$;

(b) $\lim\limits_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim\limits_{n\to\infty} \left(\frac{6^{n+1}}{(n+1)!}\cdot\frac{n!}{6^n}\right) = \lim\limits_{n\to\infty} \frac{6}{n+1} = 0 < 1$;

(c) $\lim\limits_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim\limits_{n\to\infty} \left(\frac{(n+1)^{n+1}}{(n+1)^2\cdot(n+1)!}\cdot\frac{n^2\cdot n!}{n^n}\right) = \lim\limits_{n\to\infty} \frac{n^2}{(n+1)^2}\cdot\lim\limits_{n\to\infty} \frac{(n+1)^n}{n^n} = \lim\limits_{n\to\infty} \frac{n^2}{n^2}\cdot \lim\limits_{n\to\infty} \left(1+\frac{1}{n}\right)^n = 1\cdot\mathrm{e} > 1$,

the series (a) converges; (b) converges; (c) does not converge (it diverges to $+\infty$). □

5.234. By applying the root test (Cauchy's criterion), determine whether 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) $\sum\limits_{n=1}^{\infty} \arcsin^n \frac{2n}{2^n}$

converges.

Solution. Once again we consider 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) $\lim\limits_{n\to\infty} \sqrt[n]{a_n} = \lim\limits_{n\to\infty} \frac{\left(\frac{n+1}{n}\right)^n}{\sqrt[n]{n^3}\cdot 3} = \frac{\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n}{3\left(\lim\limits_{n\to\infty}\sqrt[n]{n}\right)^3} = \frac{\mathrm{e}}{3} < 1$;

(c) $\lim\limits_{n\to\infty} \sqrt[n]{a_n} = \lim\limits_{n\to\infty} \arcsin\frac{2n}{2^n} = \arcsin 0 = 0 < 1$.

This means that all of the examined series converge. □

5.235. Determine whether 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) $\sum\limits_{n=1}^{\infty}\frac{(-3)^n}{(6+(-1)^n)^n}$

converges.

Solution. The case (a). By l'Hospital's rule, we have

$$\lim_{x\to+\infty}\frac{\ln\left(1+\frac{1}{2^x}\right)}{\frac{1}{2^x}}=\lim_{x\to+\infty}\frac{\frac{1}{1+\frac{1}{2^x}}\left(1+\frac{1}{2^x}\right)'}{\left(\frac{1}{2^x}\right)'}=\lim_{x\to+\infty}\frac{1}{1+\frac{1}{2^x}}=1,$$

hence

$$0<\ln\left(1+\frac{1}{2^n}\right)\le\frac{2}{2^n}$$

for all sufficiently large $n\in\mathbb{N}$. However, we know that the series $\sum_{n=1}^{\infty}\frac{2}{2^n}$ is convergent. So it must be that

$$\sum_{n=1}^{\infty}\ln\left(1+\frac{1}{2^n}\right)<+\infty,$$

i. e. the examined 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). Now we will use the general version of the root test

$$\limsup_{n\to\infty}\sqrt[n]{|\,a_n\,|}=\limsup_{n\to\infty}\frac{3}{6+(-1)^n}=\frac{3}{5}<1,$$

whence it follows that the series is (absolutely) convergent. □

5.236. By any means, determine whether 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. The case (a). Since we have that

$$\lim_{n\to\infty}\frac{n^2+3n-1}{(3n-2)^2}=\lim_{n\to\infty}\frac{n^2}{9n^2}=\frac{1}{9}\neq 0,$$

it immediately 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 (a necessary condition for the convergence is not satisfied).

The case (b). We have seen that when applying the ratio (or root) test, the polynomials neither in the numerator nor in the denominator affect the value of the examined limit. Let us thus consider the series

$$\sum_{n=1}^{\infty}(-1)^{n-1}\ \frac{1}{4^n}$$

for which we have

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|=\frac{1}{4}<1.$$

However, this means that the original series is also (absolutely) convergent. □

5.237. Does the following series converge?

$$\sum_{n=1}^{\infty}(-1)^{n+1}\arctan\frac{2}{\sqrt{3n}}$$

Solution. The sequence $\left\{2/\sqrt{3n}\right\}_{n\in\mathbb{N}}$ is apparently 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 we have an alternating series which satisfies 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 so-called Leibniz criterion), and this is satisfied:

$$\lim_{n\to\infty}\arctan\frac{2}{\sqrt{3n}}=\arctan 0=0,\text{ i. e. }\lim_{n\to\infty}\left((-1)^{n+1}\arctan\frac{2}{\sqrt{3n}}\right)=0.$$

□

5.238. Determine whether the series

(a) $\sum\limits_{n=1}^{\infty}\frac{\sin n}{n^2}$;

(b) $\sum\limits_{n=1}^{\infty}\frac{\cos(\pi n)}{\sqrt[3]{n^2}}$

converges absolutely, converges conditionally, or does not converge at all.

Solution. The case (a). It is easy to show that this series converges absolutely. For instance,

$$\sum_{n=1}^{\infty}\left|\frac{\sin n}{n^2}\right|\le\sum_{n=1}^{\infty}\frac{1}{n^2}<\sum_{n=0}^{\infty}\frac{1}{2^n}=2,$$

and the second inequality has already been proven.

The case (b). We can see that $\cos(\pi n)=(-1)^n$, $n\in\mathbb{N}$. So we have an alternating series such that the sequence of the absolute values of its terms is decreasing. Therefore, from the limit

$$\lim_{n\to\infty}\frac{1}{\sqrt[3]{n^2}}=0$$

it follows that the series is convergent. On the other hand,

$$\sum_{n=1}^{\infty}\left|\frac{\cos(\pi n)}{\sqrt[3]{n^2}}\right|=\sum_{n=1}^{\infty}\frac{1}{\sqrt[3]{n^2}}\ge\sum_{n=1}^{\infty}\frac{1}{n}=+\infty.$$

Thus the series converges conditionally. □

5.239. Calculate the series

(a) $\sum\limits_{n=1}^{\infty}\left(\frac{1}{\sqrt{n}}-\frac{1}{\sqrt{n+1}}\right)$;

(b) $\sum\limits_{n=0}^{\infty}\frac{5}{3^n}$;

(c) $\sum\limits_{n=1}^{\infty}\left(\frac{3}{4^{2n-1}}+\frac{2}{4^{2n}}\right)$;

(d) $\sum\limits_{n=1}^{\infty}\frac{n}{3^n}$;

(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). Apparently, this is five times the convergent geometric series with the common ratio $q = 1/3$, hence

$$\sum_{n=0}^{\infty} \tfrac{5}{3^n} = 5 \sum_{n=0}^{\infty} \left(\tfrac{1}{3}\right)^n = 5 \cdot \tfrac{1}{1-\frac{1}{3}} = \tfrac{15}{2}.$$

The case (c). We have that (substituting $m = n - 1$)

$$\sum_{n=1}^{\infty} \left(\tfrac{3}{4^{2n-1}} + \tfrac{2}{4^{2n}}\right) = \tfrac{3}{4} \sum_{n=1}^{\infty} \left(\tfrac{1}{4^{2n-2}}\right) + \tfrac{2}{16} \sum_{n=1}^{\infty} \left(\tfrac{1}{4^{2n-2}}\right) =$$
$$\left(\tfrac{3}{4} + \tfrac{2}{16}\right) \sum_{m=0}^{\infty} \tfrac{1}{4^{2m}} = \tfrac{14}{16} \sum_{m=0}^{\infty} \left(\tfrac{1}{16}\right)^m = \tfrac{14}{16} \cdot \tfrac{1}{1-\frac{1}{16}} = \tfrac{14}{15}.$$

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

$$s_n = \tfrac{1}{3} + \tfrac{2}{3^2} + \tfrac{3}{3^3} + \cdots + \tfrac{n}{3^n}, \quad n \in \mathbb{N},$$

we immediately obtain that

$$\tfrac{s_n}{3} = \tfrac{1}{3^2} + \tfrac{2}{3^3} + \cdots + \tfrac{n-1}{3^n} + \tfrac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Thus

$$s_n - \tfrac{s_n}{3} = \tfrac{1}{3} + \tfrac{1}{3^2} + \tfrac{1}{3^3} + \cdots + \tfrac{1}{3^n} - \tfrac{n}{3^{n+1}}, \quad n \in \mathbb{N}.$$

Since $\lim\limits_{n\to\infty} \frac{n}{3^{n+1}} = 0$, we get

$$\sum_{n=1}^{\infty} \tfrac{n}{3^n} = \lim_{n\to\infty} \tfrac{3}{2}\left(s_n - \tfrac{s_n}{3}\right) = \tfrac{3}{2} \lim_{n\to\infty} \sum_{k=1}^{n} \tfrac{1}{3^k} =$$
$$\tfrac{3}{2} \sum_{k=1}^{\infty} \left(\tfrac{1}{3}\right)^k = \tfrac{3}{2}\left(\tfrac{1}{1-\frac{1}{3}} - 1\right) = \tfrac{3}{4}.$$

The case (e). It suffices to use the form (the so-called partial fraction decomposition)

$$\tfrac{1}{(3n+1)(3n+4)} = \tfrac{1}{3} \cdot \tfrac{1}{3n+1} - \tfrac{1}{3} \cdot \tfrac{1}{3n+4}, \quad n \in \mathbb{N} \cup \{0\},$$

which gives

$$\sum_{n=0}^{\infty} \tfrac{1}{(3n+1)(3n+4)} = \lim_{n\to\infty} \tfrac{1}{3}\left(1 - \tfrac{1}{4} + \tfrac{1}{4} - \tfrac{1}{7} + \tfrac{1}{7} - \tfrac{1}{10} + \cdots + \tfrac{1}{3n+1} - \tfrac{1}{3n+4}\right)$$
$$= \lim_{n\to\infty} \tfrac{1}{3}\left(1 - \tfrac{1}{3n+4}\right) = \tfrac{1}{3}.$$

□

5.240. Verify that

$$\sum_{n=1}^{\infty} \tfrac{1}{n^2} < \sum_{n=0}^{\infty} \tfrac{1}{2^n}.$$

Solution. We can immediately see that

$$1 \le 1, \quad \tfrac{1}{2^2} + \tfrac{1}{3^2} < 2 \cdot \tfrac{1}{2^2} = \tfrac{1}{2}, \quad \tfrac{1}{4^2} + \tfrac{1}{5^2} + \tfrac{1}{6^2} + \tfrac{1}{7^2} < 4 \cdot \tfrac{1}{4^2} = \tfrac{1}{4},$$

or the general bound

$$\tfrac{1}{(2^n)^2} + \cdots + \tfrac{1}{(2^{n+1}-1)^2} < 2^n \cdot \tfrac{1}{(2^n)^2} = \tfrac{1}{2^n}, \quad n \in \mathbb{N}.$$

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 $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is absolutely convergent.

Let us specify that

$$\sum_{n=1}^{\infty} \tfrac{1}{n^2} = \tfrac{\pi^2}{6} < 2 = \sum_{n=0}^{\infty} \tfrac{1}{2^n}.$$

□

5.241. Examine convergence of the series

$$\sum_{n=1}^{\infty} \ln \tfrac{n+1}{n}.$$

Solution. Let us try to add up the terms of this series. We have that

$$\sum_{n=1}^{\infty} \ln \tfrac{n+1}{n} = \lim_{n\to\infty} \left(\ln \tfrac{2}{1} + \ln \tfrac{3}{2} + \ln \tfrac{4}{3} + \cdots + \ln \tfrac{n+1}{n}\right) =$$

$$\lim_{n\to\infty} \ln \tfrac{2\cdot 3\cdot 4\cdots(n+1)}{1\cdot 2\cdot 3\cdots n} = \lim_{n\to\infty} \ln(n+1) = +\infty.$$

Thus the series diverges to $+\infty$. □

5.242. Prove that the series

$$\sum_{n=0}^{\infty} \arctan \tfrac{n^2+2n+3\sqrt{n}+4}{n+1}; \qquad \sum_{n=1}^{\infty} \tfrac{3^n+1}{n^3+n^2-n}$$

do not converge.

Solution. Since

$$\lim_{n\to\infty} \arctan \tfrac{n^2+2n+3\sqrt{n}+4}{n+1} = \lim_{n\to\infty} \arctan \tfrac{n^2}{n} = \tfrac{\pi}{2}$$

and

$$\lim_{n\to\infty} \tfrac{3^n+1}{n^3+n^2-n} = \lim_{n\to\infty} \tfrac{3^n}{n^3} = +\infty,$$

the necessary condition $\lim\limits_{n\to\infty} a_n = 0$ for the series $\sum_{n=n_0}^{\infty} a_n$ to converge does not hold. □

5.243. What is the series

$$\sum_{n=2}^{\infty} \tfrac{1}{\sqrt[n]{\ln n}}?$$

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 (5.21),

$$\lim_{n\to\infty} \sqrt[n]{\ln n} = 1, \quad \text{i. e.} \quad \lim_{n\to\infty} \tfrac{1}{\sqrt[n]{\ln n}} = 1.$$

Thus the series is not convergent. Since all its terms are non-negative, it must diverge to $+\infty$. □

5.244. Find out whether the series

(a) $\sum\limits_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n}$;

(b) $\sum\limits_{n=1}^{\infty} \frac{n^2+1}{n^3}$;

(c) $\sum\limits_{n=1}^{\infty} \frac{1}{n-\ln n}$

converges.

Solution. All of the three enlisted series consist of non-negative terms only, so the series either converges, or diverges to $+\infty$. We have

(a) $\sum\limits_{n=0}^{\infty} \frac{1}{(n+1)\cdot 3^n} \le \sum\limits_{n=0}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{1-\frac{1}{3}} < +\infty$;

(b) $\sum\limits_{n=1}^{\infty} \frac{n^2+1}{n^3} \ge \sum\limits_{n=1}^{\infty} \frac{n^2}{n^3} = \sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty$;

(c) $\sum_{n=1}^{\infty} \frac{1}{n-\ln n} \geq \sum_{n=1}^{\infty} \frac{1}{n} = +\infty.$

Hence it follows that (a) converges; (b) diverges to $+\infty$; (c) diverges to $+\infty$. □

5.245. 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 $k/2$:

$$\underbrace{1+\frac{1}{2}}_{>\frac{1}{2}}+\underbrace{\frac{1}{3}+\frac{1}{4}}_{>\frac{1}{4}+\frac{1}{4}=\frac{1}{2}}+\underbrace{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}_{>\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}}+\dots$$

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 $1/2$. □

5.246. 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.

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:

$$\sum_{n=1}^{\infty} \frac{1}{(1+i)^n} = \frac{1}{1-\frac{1}{1+i}} = \frac{1+i}{i} = 1-i.$$

□

5.247. Consider a square with sides of length $a > 0$. Now consider the square whose vertices are the midpoints of the original square's sides. Then consider the square whose vertices are again the midpoints of the sides of the previous square; and so on. Determine the sum of the areas and the sum of the perimeters of all these (infinitely many) squares. ○

5.248. 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 having the radius of 2^{-n}. What is the area of an arbitrary figure consisting of all these semidiscs, supposing the semicircles do not overlap? ○

5.249. Solve the equation

$$1 - \tan x + \tan^2 x - \tan^3 x + \tan^4 x - \tan^5 x + \cdots = \tfrac{\tan 2x}{\tan 2x+1}.$$

○

5.250. Determine

$$\sum_{n=1}^{\infty} \left(\tfrac{1}{2^{n-1}} + \tfrac{2}{3^{n-1}}\right).$$

○

5.251. Calculate

$$\sum_{n=1}^{\infty} \sqrt[5n]{n^2+2n+1}.$$

○

5.252. Prove the convergence of the series

$$\sum_{n=1}^{\infty} \tfrac{3^n+2^n}{6^n}.$$

and find its value. ○

5.253. Calculate the series

(a) $\sum\limits_{n=1}^{\infty} \frac{2n-1}{2^n}$;

(b) $\sum\limits_{n=0}^{\infty} \frac{n+1}{3^n}$.

○

5.254. Sum up

$$\tfrac{1}{1\cdot 3} + \tfrac{1}{3\cdot 5} + \tfrac{1}{5\cdot 7} + \cdots = \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)(2n+1)}.$$

○

5.255. Using the partial fraction decomposition, calculate

(a) $\sum\limits_{n=2}^{\infty} \frac{1}{n^2-1}$;

(b) $\sum\limits_{n=1}^{\infty} \frac{1}{n^3+3n^2+2n}$.

○

5.256. Determine the value of the convergent series

$$\sum_{n=0}^{\infty} \frac{1}{4n^2-1}.$$

○

5.257. Calculate the series

$$\sum_{n=1}^{\infty} \frac{1}{n^2+3n}.$$

○

5.258. 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). ○

5.259. Determine whether the series

$$\sum_{n=0}^{\infty} \frac{2^n + (-2)^n}{5^n}$$

converges. ○

5.260. Prove the following statement: If a series $\sum_{n=0}^{\infty} a_n$ converges, then $\lim\limits_{n\to\infty} \sin(3a_n + \pi) = 0$.
○

5.261. For which $\alpha \in \mathbb{R}$; $\beta \in \mathbb{Z}$; $\gamma \in \mathbb{R} \setminus \{0\}$ do the series $\sum\limits_{n=120}^{\infty} \frac{e^{-\alpha n}}{n}$; $\sum\limits_{n=240}^{\infty} \frac{\beta^n \cdot n!}{n^n}$; $\sum\limits_{n=360}^{\infty} \frac{n}{\gamma^n}$ converge? ○

5.262. 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. ○

5.263. Find out whether the limit

$$\lim_{n\to\infty} \left(\tfrac{1}{n^2} + \tfrac{2}{n^2} + \cdots + \tfrac{n-1}{n^2}\right)$$

is finite. Let us warn that one cannot make use of the sums

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \qquad \sum_{n=2}^{\infty} \frac{n-1}{n^2} = +\infty.$$

○

5.264. 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. ○

5.265. Let us remind that the harmonic series diverges; i. e.

$$\sum_{n=1}^{\infty}\tfrac{1}{n}=+\infty.$$

Determine whether 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 divergent as well. ○

5.266. 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. ○

5.267. Find out whether the two series

$$\sum_{n=1}^{\infty}(-1)^n\,\tfrac{(n!)^2}{(2n)!};\qquad \sum_{n=1}^{\infty}(-1)^n\,\tfrac{n^7-n^4+n}{n^8+2n^6+n}$$

converge absolutely, converge conditionally, or do not converge at all. ○

5.268. Does the series

$$\sum_{n=1}^{\infty}(-1)^{n+1}\,\tfrac{\sqrt[3]{n}+\sqrt[5]{n}+1}{n+\sqrt[5]{n}}$$

converge? ○

5.269. 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. ○

Solutions to the exercises

5.2. $P(x) = (-\frac{3}{5} - \frac{4}{5}i)x^2 + (2 + 3i)x - \frac{3}{5} - \frac{14}{5}i.$

5.11. $3x^2 - 2x - 4.$

5.12. $\left(2x^2 - 5\right)/3$; eg. $\left(\frac{2}{3}x^2 - \frac{5}{3}\right)^3.$

5.13. $a = 1, b = -2, c = 0, d = 1.$

5.14. $x^3 + x^2 - x + 2.$

5.15. Infinitely many.

5.16. $P(x) = x^3 - 2x^2 + 5x - 3$; $Q(x) = x^3 - 2x^2 + 3x - 3.$

5.17. $x^5 - 2x^4 - 5x + 2.$

5.18. $x^2.$

5.19. $x^3 - 2x + 5$; $x^3 - x + 6.$

5.20. Infinitely many.

5.21. Eg. $x^2 - 3x + 6.$

5.22. $S_1(x) = \frac{1}{2}(x+1)^3 - \frac{3}{2}(x+1) + 1, x \in [-1, 0]$; $S_2(x) = -\frac{1}{2}x^3 + \frac{3}{2}x^2, x \in [0, 1].$

5.23. $S_1(x) = \frac{1}{2}(x+1)^3 - \frac{3}{2}(x+1) + 1, x \in [-1, 0]$; $S_2(x) = -\frac{1}{2}x^3 + \frac{3}{2}x^2, x \in [0, 1].$

5.24. $S_1(x) \equiv x$; $S_2(x) \equiv x.$

5.25. $S_1(x) \equiv 1$; $S_2(x) \equiv 1.$

5.26. $S_i(x) = x + 3, x \in [-3 + i - 1, -3 + i]$; $i \in \{1, 2\}.$

5.27. $S_1(x) = 1 - \frac{11}{20}x + \frac{1}{20}x^3$; $S_2(x) = \frac{1}{2} - \frac{2}{5}(x-1) + \frac{3}{20}(x-1)^2 - \frac{1}{40}(x-1)^3.$

5.29.

$$\begin{aligned} \sup A &= 6, & \inf A &= -3; \\ \sup B &= \frac{1}{4}, & \inf B &= -1; \\ \sup C &= 9, & \inf C &= -9. \end{aligned}$$

5.30. It can easily be shown that

$$\sup A = \frac{3}{2}, \qquad \inf A = 0.$$

5.31. Clearly

$$\inf \mathbb{N} = 1, \quad \sup \mathcal{M} = 0, \quad \inf \mathcal{J} = 0, \quad \sup \mathcal{J} = 5.$$

5.32. We can, for instance, set

$$M := \mathbb{Z} \setminus \mathbb{N}; \qquad N := \mathbb{N}.$$

5.33. Consider any singleton (one-element set) $X \subset \mathbb{R}$.

5.34. The set C must be a singleton. Thus, let us choose $C = \{0\}$, for example. Now we can take $A = (-1, 0)$, $B = (0, 1)$.

5.40. We have

$$\lim_{n\to\infty}\left(\frac{1}{n^2} + \frac{2}{n^2} + \cdots + \frac{n-2}{n^2} + \frac{n-1}{n^2}\right) = \lim_{n\to\infty}\left(\frac{1+n-1}{n^2} \cdot \frac{n-1}{2}\right) = \frac{1}{2}.$$

5.41. It can easily be shown that

$$\lim_{n\to\infty} \frac{\sqrt{n^3 - 11n^2 + 2} + \sqrt[5]{n^7 - 2n^5 - n^3} - n + \sin^2 n}{2 - \sqrt[3]{5n^4 + 2n^3 + 5}} = -\infty.$$

5.42. The limit is equal to 1.

5.43. We can, for instance, set

$$x_n := n, \quad y_n := -n + 1, \qquad n \in \mathbb{N}.$$

5.44. The answer is ± 1.

5.45. The result is

$$\limsup_{n\to\infty} a_n = 1, \qquad \liminf_{n\to\infty} a_n = 0.$$

5.46. We have

$$\liminf_{n\to\infty} \left((-1)^n \left(1 + \frac{1}{n} \right)^n + \sin \frac{n\pi}{4} \right) = -\mathrm{e} - \frac{\sqrt{2}}{2}.$$

5.62. The examined function is continuous on the whole $\mathbb{R}$.

5.63. 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.

5.64. It is necessary to set $f(0) := 0$.

5.65. The function is continuous iff $p = 2$.

5.66. The correct answer is $a = 4$.

5.67. It holds that

$$\lim_{x\to 0+} \frac{\sin^8 x}{x^3} = \lim_{x\to -\infty} \frac{\sin^8 x}{x^3} = 0.$$

5.70. The only solution is $x = -1$.

5.71. It does.

5.116. $r = +\infty$.

5.117. 1.

5.118. 3.

5.119. $[-1, 1]$.

5.120. $x \in \left[2 - \frac{1}{3}, 2 + \frac{1}{3}\right]$.

5.121. It is.

5.122.

(a) True.
(b) False.
(c) False.
(d) True.

5.123. $1 - \frac{\pi^2}{10^2 \cdot 2} + \frac{\pi^4}{10^4 \cdot 4!}$.

5.124. The error lies in the interval $(0, 1/200)$.

5.125. $\sum_{n=0}^{\infty} \frac{\mathrm{e}}{n!} (x-1)^n$; $\sum_{n=0}^{\infty} \frac{\ln^n 2}{n!} x^n$.

5.126. $f(x) = x$, $x \in \mathbb{R}$; it is.

5.127. It does not.

5.128. (a) $1 - \frac{\pi^2}{18^2 \cdot 2!} + \frac{\pi^4}{18^4 \cdot 4!}$; (b) $\frac{1}{2} - \frac{1}{5 \cdot 2^5}$.

5.129. $\sum_{n=0}^{\infty} \frac{1}{(2n+1)\, n!} x^{2n+1}$.

5.130. $a > 1$.

5.131. $[-\sqrt[3]{2}, \sqrt[3]{2})$.

5.133. $x > 2$.

5.134. The series is absolutely convergent.

5.135. $\ln(3/2)$.

5.136. $\frac{x(1-x)}{(1+x)^3}$.

5.137. (a) $\frac{1}{2}\ln\frac{1+x}{1-x}$; (b) $\frac{1+x}{(1-x)^3}$.

5.138. 2/9.

5.139. $x\,\mathrm{e}^{\frac{x^3}{2}}$.

5.144. (a) no; (b) no; (c) yes; (d) yes; (e) no; (f) no; (g) no; (h) yes.

5.145. The functions (a), (e) are odd; the functions (c), (d) are even.

5.146. It is periodic, the prime period being (a) 2π; (b) $\pi/3$.

5.147. The functions f and g are even, so it suffices to consider the graphs of the functions $y = \mathrm{e}^x$, $x \in [0, +\infty)$ and $y = \ln x$, $x \in (0, +\infty)$.

5.148. The given function is even, so to draw its graph, it suffices to know the graph of the function $y = 2^x$, $x \in (-\infty, 0]$.

5.149. $(\sinh x)' = \cosh x$; $(\cosh x)' = \sinh x$; $(\tanh x)' = \frac{1}{\cosh^2 x}$; $(\coth x)' = -\frac{1}{\sinh^2 x}$.

5.150. $\frac{1}{\sqrt{1+x^2}}$.

5.152. $x^4 + 2x^3 - x^2 + x - 2$.

5.153. $x^4 + 2x^3 - 2x^2 + x + 2$.

5.154. $x^4 + 3x^3 - 3x^2 - x - 1$.

5.155. 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 $\varepsilon \le 1$. Since if $\varepsilon > 1$, we can set $\delta = 1$.

5.156. Existence of the limit and the equality

$$\lim_{x\to -1} \frac{(1+x)^2 - 3}{2} = -\frac{3}{2}$$

follows, for example, from the choice $\delta := \varepsilon$ for $\varepsilon \in (0, 1)$.

5.157. Since $-\,(x-2)^4 < x$ for $x < 0$, we get $3\,(x-2)^4/2 > -x$ for $x < 0$.

5.158. 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.

5.159. The former limit equals $+\infty$, the latter does not exist.

5.160. The limit can be determined by a lot of means. For instance:

$$\begin{aligned}
\lim_{x\to 0} \frac{\tan x - \sin x}{\sin^3 x} &= \lim_{x\to 0} \left(\frac{\tan x - \sin x}{\sin^3 x} \cdot \frac{\cot x}{\cot x} \right) \\
&= \lim_{x\to 0} \frac{1-\cos x}{\cos x \cdot \sin^2 x} = \lim_{x\to 0} \frac{1-\cos x}{\cos x\,\left(1-\cos^2 x\right)} \\
&= \lim_{x\to 0} \frac{1}{\cos x\,(1+\cos x)} = \frac{1}{2}.
\end{aligned}$$

5.161. We have that

$$\lim_{x\to \pi/6} \frac{2\sin^3 x + 7\sin^2 x + 2\sin x - 3}{2\sin^3 x + 3\sin^2 x - 8\sin x + 3} = \lim_{x\to \pi/6} \frac{\sin x + 1}{\sin x - 1} = -3.$$

5.162. We have

$$\lim_{x\to 1} \frac{x^m - 1}{x^n - 1} = \frac{m}{n}.$$

5.163. After multiplying by the fraction

$$\frac{\sqrt{x^2+x}+x}{\sqrt{x^2+x}+x},$$

we can easily get that

$$\lim_{x\to+\infty}\left(\sqrt{x^2+x}-x\right)=\frac{1}{2}.$$

5.164. We have

$$\lim_{x\to+\infty}\left(x\sqrt{1+x^2}-x^2\right)=\frac{1}{2}.$$

5.165. We have

$$\lim_{x\to 0}\frac{\sqrt{2}-\sqrt{1+\cos x}}{\sin^2 x}=\frac{\sqrt{2}}{8}.$$

5.166. By extending the given fraction, we can obtain

$$\lim_{x\to 0}\frac{\sin(4x)}{\sqrt{x+1}-1}=8.$$

5.167. We have that

$$\lim_{x\to 0-}\frac{\sqrt{1+\tan x}-\sqrt{1-\tan x}}{\sin x}=1.$$

5.168. Apparently,

$$\lim_{x\to-\infty}\frac{2^x+\sqrt{1+x^2-x^9}-7x^5+44x^2}{3^x+\sqrt[5]{6x^6+x^2}-18x^5-592x^4}=\frac{7}{18}.$$

5.169. The statement is false. For example, consider

$$f(x):=\frac{1}{x},\quad x\in(-\infty,0);\qquad g(x):=x,\quad x\in\mathbb{R}.$$

5.170.

$$\lim_{n\to\infty}\left(\frac{n}{n+5}\right)^{2n-1}=\mathrm{e}^{-10}.$$

5.178. (a) $v(0)=6\,\mathrm{m/s}$; (b) $t=3\,\mathrm{s}$, $s(3)=16\,\mathrm{m}$; (c) $v(4)=-2\,\mathrm{m/s}$, $a(4)=-2\,\mathrm{m/s^2}$.

5.179. $f'(x_0)=\frac{1}{2\sqrt{x_0}}$.

5.180. It does not because the one-sided derivatives differ (concretely: $\pi/2$ from the right and $-\pi/2$ from the left).

5.181. It does.

5.182. It does not.

5.183. $f(x):=|x-5|+|x-9|$.

5.184. For instance, let $f=g$ take 1 at rational numbers and -1 at irrational ones.

5.185. (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) $\frac{2(1-2x)}{(1-x+x^2)^2}$.

5.186. (a) $\frac{7}{8}\,x^{-\frac{1}{8}}$; (b) $\operatorname{cosec} x=\frac{1}{\sin x}$.

5.187. $\cos x\cdot\cos(\sin x)\cdot\cos(\sin(\sin x))$.

5.188. $f'(x)=\frac{1}{\sqrt{1+2x-x^2}}+1$, $x\in\left(1-\sqrt{2},1+\sqrt{2}\right)$.

5.189. $\frac{\cos x}{3\sqrt[3]{\sin^2 x}}$.

5.190. $\frac{1+2x^2}{\sqrt{1+x^2}} + x^2\,\mathrm{e}^x$.

5.191. -8.

5.192. $\frac{2x^2}{1-x^6}\sqrt[3]{\frac{1+x^3}{1-x^3}}$.

5.193. $\ln^2\left(x+\sqrt{1+x^2}\right)$, $x\in\mathbb{R}$.

5.194. $f'(x) = -\frac{1}{x}\left(\log_x \mathrm{e}\right)^2$, $x>0$, $x\neq 1$.

5.195. $\left[f(x)g(x)h(x)k(x)\right]' = f'(x)g(x)h(x)k(x) + f(x)g'(x)h(x)k(x) + f(x)g(x)h'(x)k(x) + f(x)g(x)h(x)k'(x)$.

5.196. $\frac{x^3\,(x+1)^2\,\sqrt[3]{x+2}}{(x+3)^2}\left(\frac{3}{x}+\frac{2}{x+1}+\frac{1}{3(x+2)}-\frac{2}{x+3}\right)$.

5.207. 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 $(\sqrt{3}/8)a^2$.

5.208. 4 m × 4 m × 2 m.

5.209. $28 = 24+4$.

5.210. $a=1$.

5.211. $2\sqrt{5}\,r$.

5.212. It is the square with sides of length c).

5.213. $h=\frac{4}{3}R$, $r=\frac{2\sqrt{2}}{3}R$.

5.214. It is the equilateral triangle (with area $\sqrt{3}\,p^2/36$).

5.215. $\left[2,-1/2\right]$, $\left[-2,-1/2\right]$.

5.216. $v=2r$.

5.217. The closest point is $[1,1]$, the distance then $2\sqrt{2}$.

5.218. The closest point is $[-1,1]$, distance $3\sqrt{2}$.

5.219. $t=1,5s$, the distance will be $\sqrt{5}$ units.

5.220. It will happen at the time $t=\frac{5}{13}\,s$, the distance being $\frac{\sqrt{13}}{13}$ units.

5.221. $P=\pi rv+\pi r^2 \implies v=\frac{P-\pi r^2}{\pi r} \implies V=\frac{1}{3}r(P-\pi r^2)$. The extremum is at $r=\sqrt{\frac{P}{3\pi}}$, the substitution gives $V=\frac{2\pi}{3}$ cm^3.

5.222. (a) 12 ft/s; (b) $-59,5$ ft^2/s; (c) -1 rad/s.

5.223. At about 3 414 products per day.

5.224. Triple use of l'Hospital's rule gives

$$\lim_{x\to 0-}\frac{\sin x - x}{x^3} = -\frac{1}{6}.$$

5.225. $2/\pi$.

5.226.

$$\lim_{x\to\frac{\pi}{2}-}\left(\frac{\pi}{2}-x\right)\tan x = 1.$$

5.227.

$$\lim_{x\to+\infty}\left(\left(3^{\frac{1}{x}}-2^{\frac{1}{x}}\right)x\right) = \ln\frac{3}{2}.$$

5.228. $1/2$.

5.229. We have

$$\lim_{x\to+\infty}\left(\cos\frac{2}{x}\right)^{x^2} = \mathrm{e}^{-2}.$$

5.230. By double applying l'Hospital's rule, one obtains

$$\lim_{x\to 0}(1-\cos x)^{\sin x} = \mathrm{e}^0 = 1.$$

5.231. In both cases, the result is e^{α}.

5.232. The limit can be easily calculated by l'Hospital's rule, for instance.

5.247. $2a^2$; $4a\left(2+\sqrt{2}\right)$.

5.248. $\pi/2$.

5.249. $x=\frac{\pi}{6}+k\pi,\ x=\frac{5\pi}{6}+k\pi,\ k\in\mathbb{Z}$.

5.250. 5.

5.251. $+\infty$.

5.252. $3/2$.

5.253. (a) 3; (b) $9/4$.

5.254. $1/2$.

5.255. (a) $3/4$; (b) $1/4$.

5.256. $-1/2$.

5.257. $11/18$.

5.258. $s/2$; $3s/2$ ($s=\ln 2$).

5.259. It does.

5.260. It suffices to consider the necessary condition for convergence, namely $\lim_{n\to\infty} a_n = 0$.

5.261. $\alpha > 0$; $\beta \in \{-2,-1,0,1,2\}$; $\gamma \in (-\infty,-1)\cup(1,+\infty)$.

5.262. It is absolutely convergent.

5.263. The limit is equal to 1/2.

5.264. $A \in [0,1)$.

5.265. The value of the given series is finite – the series converges.

5.266. For example: $a_n = n/3$, $b_n = n/2$, $n \in \mathbb{N}$.

5.267. The former series converges absolutely; the latter one does conditionally.

5.268. It does.

5.269. $p \in \mathbb{R}$.

CHAPTER 6

Differential and integral calculus

we already have the menagerie, but what shall we do with it?
– we'll learn to control it...

A. Derivatives of higher orders

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 $f^{(3)}, f^{(4)}, \dots f^{(n)}$. For remembrance, we'll start with a slightly cunning problem using "only" first derivatives.

6.1. 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) $(x^n)^{(n)} = \left[(x^n)'\right]^{(n-1)} = (nx^{n-1})^{(n-1)} = \cdots = n!$.

In the previous chapter we were playing either with extremely large classes of functions — all continuous, all differentiable etc. — or only with particular functions — for example exponential, goniometric, polynomials etc. However we had only a minimum of tools and we computed everything by hand. From the qualitative point of view, we only indicated how to use the knowledge of a linear approximation of a function to its derivative to discuss the local behavior of such function near a given point. Now we will put together several results that will allow us to work with functions more easily in simulations of real problems.

By differentiation we learned how to measure instantaneous changes. In this chapter we will deal with the task of summing infinitely many of these "infinitely small" changes, e.g. how to "integrate". First though, we will clarify some things about differentiating.

In the last part of the chapter we will come back to series of functions and fill in several missing steps in our argumentation so far.

1. Differentiation

6.1. Higher order derivatives. If the first derivative $f'(x)$ of a real or a complex function 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 $f^{(2)}(x_0)$.

Function f is *double differentiable* on some interval, if it has a second derivative at each of its points. We define derivatives of higher orders inductively:

k TIMES DIFFERENTIABLE FUNCTIONS

A real or a complex function f is *differentiable* $(k+1)$ *times* at the point x_0 for some natural number k, if it is differentiable k times on some neighbourhood of the point x_0 and its k-th derivative has a derivative at the point x_0. For the k-th derivative of the function $f(x)$ we write $f^{(k)}(x)$. For $k = 0$, by 0 times differentiable functions we mean continuous functions.

If derivatives of all orders exist on an interval, we say that the function f is *smooth* on it.

For functions with continuous k-th derivative we use the denotation *the class of funcitons* $C^k(A)$ on an interval A, where k can attain values $0, 1, \dots, \infty$. Often we write only C^k, if the domain is known from the context.

(e) $(\sin x)^{(n)} = \mathrm{re}(i^n \sin x) + \mathrm{im}(i^n \cos x)$. □

6.2. 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

$$\left[\ln f(x)\right]' = \tfrac{f'(x)}{f(x)}, \quad \text{tj.} \quad f'(x) = f(x)\cdot \left[\ln f(x)\right]',$$

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

$$\left(\frac{\sqrt[4]{x-1}\cdot (x+2)^3}{\mathrm{e}^x (x+132)^2}\right)' = \frac{\sqrt[4]{x-1}\cdot (x+2)^3}{\mathrm{e}^x (x+132)^2}\cdot \left[\ln \frac{\sqrt[4]{x-1}\cdot (x+2)^3}{\mathrm{e}^x (x+132)^2}\right]'$$
$$= \frac{\sqrt[4]{x-1}\cdot (x+2)^3}{\mathrm{e}^x (x+132)^2}\cdot \left[3\ln(x+2) + \frac{1}{4}\ln(x-1) - x\,\ln \mathrm{e} - 2\ln(x+132)\right]'$$
$$= \frac{\sqrt[4]{x-1}\cdot (x+2)^3}{\mathrm{e}^x (x+132)^2}\left[\frac{3}{x+2} + \frac{1}{4\,(x-1)} - 1 - \frac{2}{x+132}\right].$$

□

6.3. Let $n \in \mathbb{N}$ be arbitrary. Find the n-th derivative of function

$$y = \ln \tfrac{1+x}{1-x}, \quad x \in (-1, 1).$$

Solution. With respect of the equality

$$\ln \tfrac{1+x}{1-x} = \ln(1+x) - \ln(1-x), \quad x \in (-1,1),$$

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

$$f'(x) = \tfrac{a}{ax+1},$$
$$f''(x) = \tfrac{-a^2}{(ax+1)^2},$$
$$f^{(3)}(x) = \tfrac{2a^3}{(ax+1)^3},$$
$$f^{(4)}(x) = \tfrac{-6a^4}{(ax+1)^4}.$$

Based on these results we can figure out that

$$(6.1) \qquad f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!\,a^n}{(ax+1)^n}, \quad x \in (-1,1),\ n \in \mathbb{N}.$$

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

$$f^{(k+1)}(x) = \left(\tfrac{(-1)^{k-1}(k-1)!\,a^k}{(ax+1)^k}\right)' = \tfrac{(-1)^{k-1}(k-1)!\,a^k\,(-k)\,a}{(ax+1)^{k+1}} = \tfrac{(-1)^k k!\,a^{k+1}}{(ax+1)^{k+1}},$$

vzorec (‖6.1‖) it holds for all $n \in \mathbb{N}$. Then

$$\ln^{(n)}(1+x) = \tfrac{(-1)^{n-1}(n-1)!}{(x+1)^n}, \quad \ln^{(n)}(1-x) = -\tfrac{(n-1)!}{(-x+1)^n}, \quad x \in (-1,1).$$

From here we obtain the result

$$\left(\ln \tfrac{1+x}{1-x}\right)^{(n)} = (n-1)!\left(\tfrac{1}{(1-x)^n} - \tfrac{(-1)^n}{(1+x)^n}\right)$$

We can illustrate the concept of higher order derivatives on polynomials. Because a derivative of a polynomial is a polynomial with a degree one less than the original one, after a finite number of differentiations we get the zero polynomial. More precisely, after exactly $k+1$ differentiations, where k is the degree of the polynomial, we get zero. Of course then derivatives of all orders exist, e.g. $f \in C^\infty(\mathbb{R})$.

In the spline construction, see 5.9, we took care that the resulting functions would belong to the class $C^2(\mathbb{R})$. Their third derivatives will be sequentially constant functions. That is why the splines won't belong to $C^3(\mathbb{R})$, even though all their higher order derivatives will be 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 simple 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:*

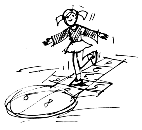
$$(f\cdot g)^{(k)}(x_0) = \sum_{i=0}^{k} \binom{k}{i} f^{(i)}(x_0) g^{(k-i)}(x_0).$$

Proof. For $k = 0$ the statement is trivial, for $k = 1$ it's Leibniz's product rule. If the equality holds for some k, by differentiating the right hand side and using Leibniz's rule we obtain a smiliar expression

$$\sum_{i=0}^{k}\binom{k}{i}\Big(f^{(i+1)}(x_0)g^{(k-i)}(x_0) + f^{(i)}(x_0)g^{(k-i+1)}(x_0)\Big).$$

In this new sum, the sum of orders of 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 $\binom{k}{j-1} + \binom{k}{j} = \binom{k+1}{j}$. □

6.2. Multiple roots and inversions of polynomials. We already computed the derivatives of polynomials in the paragraph 5.6 and it can be seen that these are smooth functions. In this case differentiation can be viewed as an injective algebraic map. Let's see how we can use differentiation for discussing multiple roots of polynomials.



First we formulate *The fundamental theorem of algebra*, whose proof will be left over to **??**.

Theorem. *Each nonzero complex polynomial $f : \mathbb{C} \to \mathbb{C}$ of degree at least one has a root.*

Thus a polynomial of degree $k > 0$ has exactly k complex roots (counting multiplicities) and can be written uniquely in the form

$$f(x) = (x-a_1)^{c_1}\cdot (x-a_q)^{c_q},$$

where $a_1, \dots, a_q$ are all roots of the polynomial f and

$$1 \le c_1, \dots, c_q \le k$$

are their multiplicities (i.e. natural numbers).

By differentiation of $f(x)$ as a function of one real variable x we get

$$f'(x) = c_1(x-a_1)^{c_1-1}\dots(x-a_q)^{c_q} + \dots$$
$$+ c_q(x-a_1)^{c_1}\dots(x-a_q)^{c_q-1}.$$

pro $x \in (-1, 1)$ a $n \in \mathbb{N}$. □

6.4. Determine the second derivative of function $y = \operatorname{tg} x$ on its whole domain, i.e. for $\cos x \neq 0$. ○

6.5. Determine the fifth and the sixth derivative of the polynomial

$$p\,(x) = \left(3x^2 + 2x + 1\right) \cdot (2x - 6) \cdot \left(2x^2 - 5x + 9\right), \quad x \in \mathbb{R}.$$

○

6.6. With no computation involved, determine the 12th derivative of function

$$y = \mathrm{e}^{2x} + \cos x + x^{10} - 5x^7 + 6x^3 - 11x + 3, \quad x \in \mathbb{R}.$$

○

6.7. Write the 26th derivative of function

$f(x) = \sin x + x^{23} - x^{18} + 15x^{11} - 13x^8 - 5x^4 - 11x^3 + 16 + \mathrm{e}^{2x}$

pro $x \in \mathbb{R}$. ○

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.

6.8. Jensen inequality. For a strictly convex function f on interval I and for arbitrary points $x_1, \dots, x_n \in I$ and real numbers $c_1, \dots, c_n > 0$ 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. Proof can be found for example in ‖**??**‖ □

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.

6.9. Prove that among all (convex) n-gons inscribed into a circle, the regular n-gon has the biggest 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

$$S_i = x_i\, h_i = r^2 \sin \tfrac{\varphi_i}{2} \cos \tfrac{\varphi_i}{2} = \tfrac{1}{2}\, r^2 \sin \varphi_i, \quad i \in \{1, \dots, n\}.$$

This implies that the area of the whole n-gon is

$$S = \sum_{i=1}^{n} S_i = \tfrac{1}{2}\, r^2 \sum_{i=1}^{n} \sin \varphi_i.$$

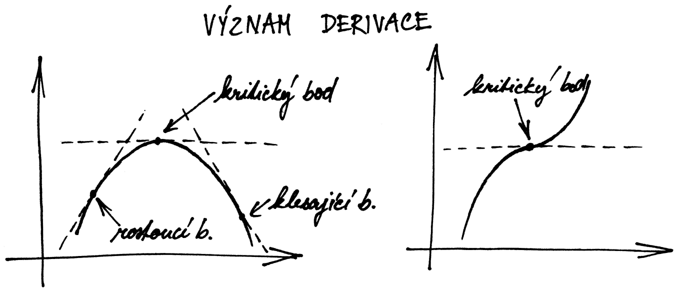
Thus we want to maximize the sum $\sum_{i=1}^{n} \sin \varphi_i$, while for values $\varphi_i \in (0, \pi)$ we clearly have

$$\varphi_1 + \cdots + \varphi_n = \sum_{i=1}^{n} \varphi_i = 2\pi. \tag{6.2}$$

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

If $c_1 = 1$ and the root a_1 is real, the value of the derivative f' at the point a_1 will be nonzero, because the first term is nonzero, while all the others will vanish after setting $x = a_1$. Similarly for other roots. Thus we verified a convenient property that a real root a of a polynomial f is multiple if and only if it is a root of its derivative f'. (We will extend this statement to all complex roots in time.)

6.3. 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 and that the sign of 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 *critical points* or *stationary points* of the given function.



If x_0 is a stationary point of function f, the behavior of function f in the neighbourhood of x_0 can vary. It can be seen for example from the behavior of function $f(x) = x^n$ in the neighbourhood of zero for arbitrary n. For odd $n > 0$, $f(x)$ will be increasing, while for even n it will be decreasing on the left side and increasing on the right side, therefore at x_0 it will attain its minimal value among points from (sufficiently small) neighbourhood of $x_0 = 0$.

We can apply this point of view to function f'. If the second derivative is nonzero, its sign determines the behavior of the first derivative. That's why at the critical point x_0 the derivative $f'(x)$ will be increasing if the second derivative is positive and decreasing if the second derivative is negative. If it's increasing though, it means that it will necessarily be negative to the left of the critical point and positive to the right of it. In that case, function f is decreasing to the left of the critical point and increasing to the right of it. That means f attains its minimal value among all points from (sufficiently small) neighbourhood of x_0 at the point x_0.

On the other hand, if the second derivative is negative at x_0, the first derivative is decreasing, thus negative to the left of x_0 and positive to the right of it. Function f will then attain its maximal value among all values from some neighbourhood.

A function that is differentiable on (a, b) and continuous on $[a, b]$ certainly has an absolute maximum and minimum of this interval. It can be attained only at its boundary or at a point with zero derivative, i.e. in a critical point. That means critical points may be sufficient for finding extremes and second derivatives will help us determine the types of the extremes, if they are nonzero. For more precise discussion though we need better approximation of the studied function than a linear one. That's why we'll first study notions in this direction and later come back to discussing the course of functions.

interval. Then according to Jensen's inequality for $c_i = 1/n$ and $x_i = \varphi_i$, we have

$$-\sin\left(\sum_{i=1}^{n} \tfrac{1}{n}\,\varphi_i\right) \le -\sum_{i=1}^{n} \tfrac{1}{n}\,\sin\varphi_i, \quad \text{tj.} \quad \sin\left(\sum_{i=1}^{n} \tfrac{1}{n}\,\varphi_i\right) \ge \sum_{i=1}^{n} \tfrac{1}{n}\,\sin\varphi_i.$$

Moreover, we know the equality occurs exactly for $\varphi_1 = \cdots = \varphi_n$. If we express (using (‖6.2‖))

$$S = \tfrac{r^2 n}{2}\sum_{i=1}^{n} \tfrac{1}{n}\,\sin\varphi_i \le \tfrac{r^2 n}{2}\,\sin\left(\sum_{i=1}^{n} \tfrac{1}{n}\,\varphi_i\right) = \tfrac{r^2 n}{2}\,\sin\tfrac{2\pi}{n},$$

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 $\varphi_1 = \cdots = \varphi_n = 2\pi/n$. □

6.10. Isoperimitric quotient. For a closed curve in plane enclosing a planar region, we define its isoperimetric quotient as the number

$$IQ := \frac{S}{\pi(\frac{o}{2\pi})^2} = \frac{4\pi S}{o^2},$$

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, $IQ \le 1$, 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,

$$h = \cos\varphi = \cos\tfrac{\pi}{n}, \qquad \tfrac{x}{2} = \sin\varphi = \sin\tfrac{\pi}{n},$$

which yields

$$o_n = n \cdot x = 2n\sin\tfrac{\pi}{n}$$

and

$$S_n = n \cdot \tfrac{1}{2}\,hx = n\cos\tfrac{\pi}{n}\,\sin\tfrac{\pi}{n}.$$

Thus for a regular n-gon, we have

$$IQ = \frac{4\pi n\cos\frac{\pi}{n}\,\sin\frac{\pi}{n}}{4n^2\sin^2\frac{\pi}{n}} = \tfrac{\pi}{n}\,\mathrm{cotg}\,\tfrac{\pi}{n},$$

which we can verify for example for a square ($n = 4$) with a side of length a, where

$$IQ = \frac{4\pi a^2}{(4a)^2} = \tfrac{\pi}{4} = \tfrac{\pi}{4}\,\mathrm{cotg}\,\tfrac{\pi}{4}.$$

Using the limit transition for $n \to \infty$ and the limit

$$\lim_{x\to 0}\frac{\sin x}{x} = 1,$$

we get the isoperimetric quotient for a circle:

6.4. Taylor expansion. As a surprisingly easy use of Rolle's theorem we will now derive an extremely important result. It's called *Taylor expansion with a remainder.* Intuitively we can get to it by reversing our notions about power series. If we have a power series centered in a,

$$S(x) = \sum_{n=0}^{\infty} a_n(x-a)^n,$$

and we differentiate it repeatedly, we are getting power series (we know that we can differentiate such expression term after term, even if we haven't proved it yet)

$$S^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)a_n(x-a)^{n-k}.$$

At the point $x = a$ we then have $S^{(k)}(a) = k!a_k$. Then we can conversely read the last statement as an equation for a_k and rewrite the original series as

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{k!}S^{(k)}(a)(x-a)^n.$$

If we have some sufficiently smooth function $f(x)$ instead of a power series, it's suitable to ask if it can be expressed as a power series and how fast will the partial sums (i.e. approximations of function f by polynomials) converge. Our notion just suggested we can expect a good approximation by polynomials in the neighbourhood of point a.

Taylor polynomials of function f

Fot k times differentiable function f we define its *Taylor polynomial of k–th degree* by the relation

$$T_{k,a}f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f^{(3)}(a)(x-a)^3 + \cdots + \frac{1}{k!}f^{(k)}(a)(x-a)^k.$$

The precise answer looks similar to the mean value theorem, but we work with higher degrees of polynomials:

Theorem (Taylor expansion with a remainder)**.** *Let $f(x)$ be a function that is k times differentiable on interval (a, b) and continuous on $[a, b]$. Then for all $x \in (a, b)$ there exists a number $c \in (a, x)$ such that*

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \dots \\ &\quad + \frac{1}{(k-1)!}f^{(k-1)}(a)(x-a)^{k-1} + \frac{1}{k!}f^{(k)}(c)(x-a)^k \\ &= T_{k-1,a}f(x) + \frac{1}{k!}f^{(k)}(c)(x-a)^k. \end{aligned}$$

Proof. Define the remainder R (i.e. the error of the approximation for fixed x) as follows

$$f(x) = T_{k-1,a}f(x) + R$$

i.e. $R = \frac{1}{k!}r(x-a)^k$ for a suitable number r (dependant on x). Now consider function $F(\xi)$ defined by

$$F(\xi) = \sum_{j=0}^{k-1} \frac{1}{j!}f^{(j)}(\xi)(x-\xi)^j + \frac{1}{k!}r(x-\xi)^k.$$

$$IQ = \lim_{n\to\infty} \tfrac{\pi}{n} \operatorname{cotg} \tfrac{\pi}{n} = \lim_{n\to\infty} \frac{\cos\frac{\pi}{n}}{\frac{\sin\frac{\pi}{n}}{\frac{\pi}{n}}} = \tfrac{\cos 0}{1} = 1.$$

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

$$IQ = \tfrac{4\pi S}{o^2} = \tfrac{4\pi \frac{\varphi r^2}{2}}{(2r+r\varphi)^2} = \tfrac{2\pi\varphi}{(2+\varphi)^2}.$$

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

$$f'(\varphi) = 2\pi \tfrac{(2+\varphi)^2 - 2\varphi(2+\varphi)}{(2+\varphi)^4} = 2\pi \tfrac{2-\varphi}{(2+\varphi)^3}, \quad \varphi \in (0, 2\pi)$$

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.○ □

6.11. 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

$$\tfrac{1}{\lambda_i} := \tfrac{4\pi S_i}{o_i^2}, \qquad i \in \{1, \dots, n\},$$

where S_i is the area and o_i the perimeter of the i-th figure. We'll also use the denotation

$$\Lambda := \sum_{i=1}^{n} \lambda_i.$$

Recall that the isoperimetric quotient is given only by the shape of the figure and doesn't depend on its size. In particular, the value 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

$$S_i = \tfrac{o_i^2}{4\pi\lambda_i}, \qquad i \in \{1, \dots, n\},$$

we need to minimize the expression

$$S := \tfrac{1}{4\pi} \sum_{i=1}^{n} \tfrac{o_i^2}{\lambda_i}.$$

Its derivative (when x is considered as a constant parameter) is

$$\begin{aligned} F'(\xi) &= f'(\xi) + \\ &\sum_{j=1}^{k-1}\left(\frac{1}{j!}f^{(j+1)}(\xi)(x-\xi)^j - \frac{1}{(j-1)!}f^{(j)}(\xi)(x-\xi)^{j-1}\right) \\ &\quad - \frac{1}{(k-1)!}r(x-\xi)^{k-1} \\ &= \frac{1}{(k-1)!}f^{(k)}(\xi)(x-\xi)^{k-1} - \frac{1}{(k-1)!}r(x-\xi)^{k-1} \\ &= \frac{1}{(k-1)!}(x-\xi)^{k-1}(f^{(k)}(\xi) - r), \end{aligned}$$

because the expressions in the sum cancel each other out sequentially. Now it suffices to notice that $F(a) = F(x) = f(x)$ (recall that x is an arbitrarily chosen but fixed value from interval (a, b)). Then according to Rolle's theorem there exists a number c, $a < c < x$ such that $F'(c) = 0$. That's exactly the desired relation. □

6.5. Estimations for expansions with a remainder. An especially simple case of Taylor expansion is the expansion of an arbitrary polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \dots + a_1 x + a_0, \quad a_n \neq 0.$$

Because the $(n+1)$–th derivative f is identically zero, the Taylor polynomial of degree n has a zero remainder, therefore for each$x_0 \in \mathbb{R}$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \dots + \frac{1}{n!}f^{(n)}(x_0)(x - x_0)^n$$

and 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 series with a remainder. We know in advance that the remainder can be estimated by the size of the derivative which is identically zero from some degree on for polynomials.

More universally the estimation of the size of the k–th derivative on some interval can be used to estimate the error on the same interval. A special case is also the mean value theorem as an approximation by Taylor series of degree zero, see (5.9). A good examples of an expansion of an arbitrary degree are the goniometric functions sin and cos. By iterating the differentiation of function $\sin x$ we always get either sine or cosine with some sign, but in the absolute value the values won't exceed one. Thus we get a direct estimation of the speed of convergence of the power series

$$|\sin x - (T_{k,0}\sin)(x)| \leq \frac{|x|^{k+1}}{(k+1)!}.$$

It show that for x drasticall lesser than k the error will be small, but for x comparable with k or bigger it will be large. Compare with the figure of the approximation of function $\cos x$ by Taylor polynomial of degree 68 in paragraph 5.50.

As we 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 formulation of the converse implication, i.e. when the given function $f(x)$ is actually a power series.

Using Jensen's inequality for the strictly convex function $y = x^2$ (on the whole real axis), we obtain

$$\left(\sum_{i=1}^{n} c_i\, x_i\right)^2 \le \sum_{i=1}^{n} c_i\, x_i^2$$

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

$$c_i = \tfrac{\lambda_i}{\Lambda}, \quad x_i = \tfrac{o_i}{\lambda_i}, \qquad i \in \{1, \dots, n\},$$

we then get

$$\left(\sum_{i=1}^{n} \tfrac{\lambda_i}{\Lambda}\tfrac{o_i}{\lambda_i}\right)^2 \le \sum_{i=1}^{n} \tfrac{\lambda_i}{\Lambda}\left(\tfrac{o_i}{\lambda_i}\right)^2 .$$

By several simplifications, we obtain the inequality

$$\tfrac{1}{\Lambda^2}\left(\sum_{i=1}^{n} o_i\right)^2 \le \tfrac{1}{\Lambda}\sum_{i=1}^{n} \tfrac{o_i^2}{\lambda_i}$$

and then (notice that $\sum_{i=1}^{n} o_i = l$)

$$\tfrac{l^2}{\Lambda} \le \sum_{i=1}^{n} \tfrac{o_i^2}{\lambda_i},$$

with equality again occuring for

$$x_1 = \cdots = x_n, \qquad \text{tj.} \qquad \frac{o_1}{\lambda_1} = \cdots = \frac{o_n}{\lambda_n}. \tag{6.3}$$

This implies that S the smallest, if and only if (‖6.3‖) 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 (‖6.3‖) 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.

$$o_i = \tfrac{\lambda_i}{\Lambda}\, l, \qquad i \in \{1, \dots, n\}.$$

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)

$$\lambda_1 = \tfrac{4}{\pi}, \quad \lambda_2 = 1, \qquad \text{tj.} \qquad \Lambda = \lambda_1 + \lambda_2 = \tfrac{4+\pi}{\pi}.$$

Then the lengths of the respective parts are (in metres)

$$o_1 = \frac{\frac{4}{\pi}}{\frac{4+\pi}{\pi}} \cdot 1 = \tfrac{4}{4+\pi} \doteq 0,56, \quad o_2 = \frac{1}{\frac{4+\pi}{\pi}} \cdot 1 = \tfrac{\pi}{4+\pi} \doteq 0,44.$$

The area of a square with perimeter 0, 56 m (with a side of length $a = 0,14$ m) is $0,019\,6\ \mathrm{m}^2$ and the area of a circle with perimeter 0, 44 m (and radius $r \doteq 0,07$ m) is approximately $0,015\,4\ \mathrm{m}^2$. We can verify that (in m^2

$$\tfrac{l^2}{4\pi\Lambda} = \tfrac{1}{4(4+\pi)} \doteq 0,035 = 0,019\,6 + 0,015\,4.$$

□

Taylor expansions. We necessarily need the derivatives of higher orders to determine the Taylor expansion of a given function.

Corollary (Taylor's theorem)**.** *Assume that the function $f(x)$ is smooth on the interval $(a-b, a+b)$ and all of its derivatives are bounded uniformally here by a constant $M > 0$, i.e.*

$$|f^{(k)}(x)| \le M, \quad k = 0, 1, \dots, \ x \in (a-b, a+b).$$

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 the function $f(x)$.

Proof. The proof is identical with the notion in the specific case of function $\cos x$ higher. Think through the details! □

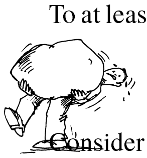
6.6. Analytic and smooth functions. If f is smooth at a, we can write a formal power series

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{k!} f^{(k)}(a)(x-a)^n.$$

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 the point a. A function is analytic on an interval, if it's analytic at its every point.

Not all smooth functions are analytic though. In fact it can be proven that for every sequence of numbers a_n we can find a smooth function, whose derivatives of order k will be these numbers a_k.[1]

To at least imagine the essence of the problem, we'll introduce (as will be seen later, a very useful) function, that has all derivatives zero at zero but is nonzero at every other point.



Consider a function defined by

$$f(x) = e^{-1/x^2}.$$

Obviously it's a well defined smooth function at all points $x \ne 0$. We'll check that at the point $x = 0$ the limit exists: $\lim_{x\to 0} f(x) = 0$. Thus we can aditionally define $f(0) = 0$ and obtain a smooth function.

By a direct computation with usage of L'Hospital's rule we'll compute the derivative and it suffices to consider only the right derivative, because 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 the function $f(x)$ at an arbitrary point $x \ne 0$ we'll get $f'(x) = e^{-1/x^2} \cdot 2x^{-3}$ a by repeated differentiating of the results we'll always get 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.

So we'll assume we've already proven that the derivative of order k of our function $f(x)$ exists and is zero at zero. While computing the following derivatives we'll proceed the same way as in the case $k = 0$ higher. We'll compute the limit of the expression $f^{(k)}(x)/x$ for $x \to 0^+$, i.e. 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 we can use L'Hospital's rule repeatedly on them. Obviously after several differentiations of both the numerator and denominator (and a similar adjustment as higher) there will be still the same expression in the denominator, while in the numerator the power will be nonnegative. Thus the whole expression necessarily has a

[1] It's a special case of so called Whitney's theorem, see. COMPLETE THE CITATION AND INFORMATION.

6.12. 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 functionsin(x) at point 0 is

$$T_0^3(\sin(x)) = x - \frac{1}{6}x^3.$$

(ii) Again $f(1) = e$,

$$\begin{aligned} f'(1) &= \left.\frac{e^x}{x} - \frac{e^x}{x^2}\right|_{x=1} = 0 \\ f^{(2)}(1) &= \left.\frac{e^x}{x} - 2\frac{e^{x\,2}}{x} + \frac{2e^x}{x^3}\right|_{x=1} = e \\ f^{(3)}(1) &= \left.\frac{e^x}{x} - 3\frac{e^{x\,2}}{x} + \frac{6e^x}{x^3} - \frac{6e^x}{x^4}\right|_{x=1} = -2e \end{aligned}$$

Thus we get the Taylor expansion of third order of function $\frac{e^x}{x}$ at point 1:

$$T_1^3(\frac{e^x}{x}) = e + \frac{e}{2}(x-1)^2 - \frac{e}{3}(x-1)^3 = e(-\frac{x^3}{3} + \frac{3x^2}{2} - 2x + \frac{5}{6}).$$

□

6.13. Determine the Taylor polynomial T_0^6 of function sin and using theorem (6.4), 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 6.4, we then estimate the size of the remainder (error) R. According to the theorem, there exists $c \in (0, \frac{\pi}{4})$ such that

$$R(\pi/4) = \left|\frac{-\cos(c)\pi^7}{7!4^7}\right| < \frac{1}{7!} \doteq 0,0002.$$

□

6.14. Find the Taylor polynomial of third order of function

$$y = \operatorname{arctg} x, \quad x \in \mathbb{R}$$

at point $x_0 = 1$. ○

6.15. Determine the Taylor expansion of third order at point $x_0 = 0$ of function

(a) $y = \frac{1}{\cos x}$;
(b) $y = \mathrm{e}^{-\frac{x^2}{2}}$;
(c) $y = \sin(\sin x)$;
(d) $y = \operatorname{tg} x$;
(e) $y = \mathrm{e}^x \sin x$

zero limit at zero, just like we've computed in the case of the first derivative higher. The same will hold true for a finite sum of such expressions, so we've found out that each derivative $f^{(k)}(x)$ at zero will exist and its value will be zero.

We've shown that our function $f(x)$ is smooth on whole $\mathbb{R}$, it's of course a nonzero function everywhere except for $x = 0$, but all its derivatives at this point are zero. Of course, then it's not an analytic function at the point $x_0 = 0$.

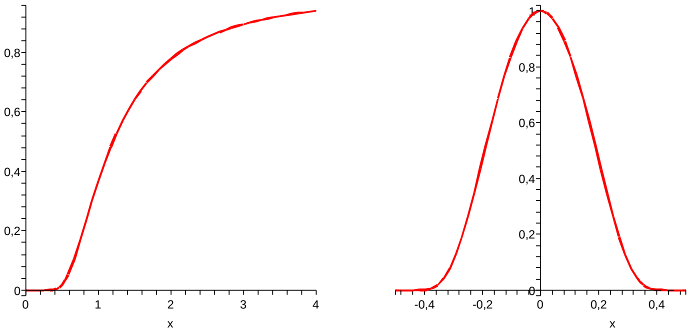
6.7. Examples of nonanalytic smooth functions. We can easily modify our function $f(x)$ from the previous paragraph in this way:

$$g(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \mathrm{e}^{-1/x^2} & \text{if } x > 0 \end{cases}.$$

Again it's a smooth function on whole $\mathbb{R}$. By another modification we can obtain a function that is nonzero in all inner points of the interval $[-a, a]$, $a > 0$ and zero elsewhere:

$$h(x) = \begin{cases} 0 & \text{if } |x| \ge a \\ \mathrm{e}^{\frac{1}{x^2-a^2} + \frac{1}{a^2}} & \text{if } |x| < a\,. \end{cases}$$

This function is again smooth on whole $\mathbb{R}$. The last two functions are on the figures, on right the parameter $a = 1$ is used.



Finally we'll show how to get smooth analogies of Heaviside functions. For two fixed real numbers $a < b$ we define the function $f(x)$ with usage of earlier defined function g in this way:

$$f(x) = \frac{g(x-a)}{g(x-a) + g(b-x)}.$$

Obviously for all $x \in \mathbb{R}$ the denominator of the fraction is positive (because for each of the intervals determined by numbers a and b at least one of the summands of the denominator is nonzero, therefore the whole denominator is positive. Thus from our definition we get a smooth function $f(x)$ on whole $\mathbb{R}$. For $x \le a$ the denominator of the fraction is zero according to the definition of g though, for $x \ge b$ the numerator and denominator are equal. On the next two figures there are these functions $f(x)$ with parameters $a = 1 - \alpha$, $b = 1 + \alpha$, where on the left we have $\alpha = 0.8$ and on the right we have $\alpha = 0.4$.

defined in a certain neighbourhood of point x_0. ○

6.16. Determine the Taylor expansion of fourth order of function $y = \ln x^2$, $x \in (0, 2)$ at point $x_0 = 1$. ○

6.17. Find the estimation of the error of the approximation

$$\ln(1+x) \approx x - \tfrac{x^2}{2}$$

for $x \in (-1, 0)$. ○

6.18. 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

$$\lim_{x\to 0+} \tfrac{x\ \sin x - x^2}{x^4}.$$

○

6.19. Determine the Taylor polynomial centered at the origin of degree at least 8 of function $y = \mathrm{e}^{2x}$, $x \in \mathbb{R}$. ○

6.20. Express the polynomial $x^3 - 2x + 5$ as a polynomial in variable $u = x - 1$. ○

6.21. Espand 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 $\ln(x+1) = \ln 1 + \sum\limits_{n=1}^{\infty} a_n x^n$, where

$$a_n = \frac{(-1)^{n+1}(n-1)!}{n!} = \frac{(-1)^{n+1}}{n}.$$

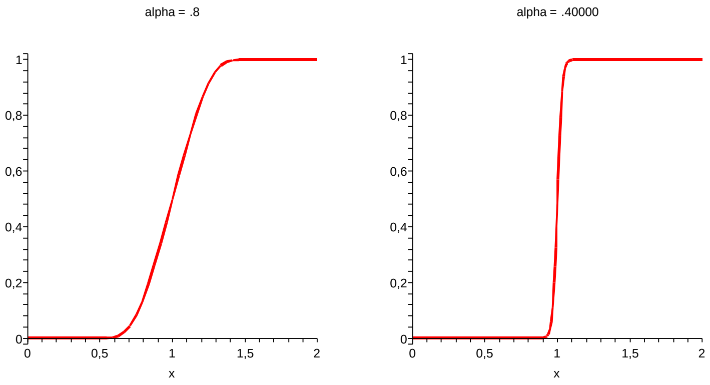
Thus we can write

$$\begin{aligned}\ln(x+1) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots \\ &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n.\end{aligned}$$

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 $x \in (-1, 1]$.



Now we can also easily create a smooth analogy of the characteristic function of the interval $[c, d]$.

Denote the higher specified function $f(x)$ with parameters $a = -\varepsilon$, $b = +\varepsilon$ as $f_\varepsilon(x)$. Now for the interval (c, d) with length $d - c > 2\varepsilon$ we 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 $(d + \varepsilon, \infty)$ and identically equal one on the interval $(c + \varepsilon, d - \varepsilon)$, moreover it's smooth everywhere and locally it's either constant or monotonic. The smaller the $\varepsilon > 0$, the faster our function jumps from zero to one around the beginning of the interval or back at the end of it.

Thus we can see that smooth functions are very "plastic" — from a local behaviour around one point we cannot deduce anything at all about the global behavior of such function. Conversely, analytic functions are completely determined just by derivatives at one point. In particular they are completely determined by their behavior on an arbitrarily small neighbourhood of a single point from their domain. In this sense, they are very "rigid".

6.8. Local behavior of functions. We've seen that the sign of the first derivative of a differentiable function determines whether it's increasing or decreasing on some neighbourhood of the given point. If derivative is zero though, it doesn't tell us much about the behavior of the function by itself.

We've already encountered the importance of the second derivative while describing critical points. Now we'll generalize the discussion of critical points for all orders. We'll start with discussing the local extremes of functions, i.e. values, that are strictly bigger or strictly smaller than all the other values from some neighbourhood of a given point.

In the following we'll consider functions with sufficiently high number of continuous derivatives, without specifically pointing this assumption out.

We say the point a in domain of f is *a critical point of order k* iff

$$f'(a) = \cdots = f^{(k)}(a) = 0, \quad f^{(k+1)}(a) \neq 0.$$

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, a Taylor expansion with a remainder gives us

$$f(x) = f(a) + \frac{1}{(k+1)!} f^{(k+1)}(c)(x-a)^{k+1}$$

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 behavior of the function $(x - a)^{k+1}$. Moreover, if $k+1$ is an even number, then the values of $f(x)$ in such neighbourhood are necessarily bigger than the value

Analogously, for the expansion at point 1, by computing the above derivatives from ‖6.21‖, we get

$$\begin{aligned}\ln(x+1) &= \ln(2) + \frac{1}{2}(x-1) - \frac{1}{8}(x-1)^2 + \frac{1}{3\cdot 2^3}(x-1)^3 - \dots \\ &= \ln(2) + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n\cdot 2^n}(x-1)^n,\end{aligned}$$

and for the radius of convergence of this series, we get

$$r = \frac{1}{\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|} = \frac{1}{\lim_{n\to\infty}\frac{\frac{1}{2^{n+1}(n+1)}}{\frac{1}{2^n n}}} = 1.$$

The first series converges for $-1 < x \leq 1$, the second for $-1 < x \leq 3$. □

6.22. Expand the function

(a) $y = \ln\frac{1+x}{1-x}, \quad x \in (-1,1);$

(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

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n, \quad x \in (-1,1),$$

i.e.

$$\ln(1-x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (-x)^n = -\sum_{n=1}^{\infty} \frac{1}{n} x^n, \quad x \in (-1,1).$$

In total, we have

$$\ln\frac{1+x}{1-x} = \ln(1+x) - \ln(1-x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}+1}{n} x^n = \sum_{n=1}^{\infty} \frac{2}{2n-1} x^{2n-1}$$

for $x \in (-1,1)$.

Case (b). Similarly, the well known identity

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n, \quad x \in \mathbb{R},$$

implies

$$e^{x^2} = \sum_{n=0}^{\infty} \frac{1}{n!}\left(x^2\right)^n = \sum_{n=0}^{\infty} \frac{1}{n!} x^{2n}, \quad x \in \mathbb{R},$$

and

$$x^2 e^{-2x} = x^2 \sum_{n=0}^{\infty} \frac{1}{n!}(-2x)^n = \sum_{n=0}^{\infty} \frac{(-2)^n}{n!} x^{n+2}, \quad x \in \mathbb{R}.$$

Hence

$$e^{x^2} + x^2 e^{-2x} = \sum_{n=0}^{\infty} \frac{x^{2n} + (-2)^n x^{n+2}}{n!}, \quad x \in \mathbb{R}.$$

□ R

$f(a)$ and obvisouly the point a is a point of local minimum then. If k is even though, then the values on left are small and on right bigger than $f(a)$, so an extreme doesn't occur even locally. On the other we can notice that the graph of function $f(x)$ intersects its tangent $y = f(a)$ at point $[a, f(a)]$.

Conversely, if $f^{(k+1)}(a) < 0$, then because of the same reasoning it's a local maximum for odd k a again the extreme doesn't occur for even k.

6.9. Convex and concave functions. We say that differentiable function f is *concave* at point a, if in a certain neighbourhood its graph lies completely below the tangent at point $[a, f(a)]$, i.e. we require

$$f(x) \leq f(a) + f'(a)(x-a).$$

Conversely, we say that f is *convex* at point a, if its graph is above the tangent at point a, i.e.

$$f(x) \geq f(a) + f'(a)(x-a).$$

A function is convex or concave on an interval, if it has this property in its every point.

Moreover suppose that function f has continuous second derivatives in a neighbourhood of point a. From the Taylor expansion of second order with a remainder we obtain

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(c)(x-a)^2.$$

Then obviously the function is convex, whenever $f''(a) > 0$, and concave, whenever $f''(a) < 0$.

If the second derivative is zero, we cab zse derivatives of higher orders. 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, clearly 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.

6.10. Inflection points. Point a is called *an inflective point* of a differentiable function f, if the graph of function $f$crosses from one side of the tangent to the other.

Suppose f has continuous third derivatives and write the Taylor expansion of third order with a 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 nonzero point of the second derivative such that $f'''(a) \neq 0$, then the first derivative is nonzero on some neighbourhood as well and clearly it's an inflective point. 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 nonzero point of the second derivative and simultaneously an inflective point, then clearly on some small neighbourhood of a the function is concave on one side and convex on the other. Thus we can also see the inflective points as the points of the crossover betwenn concave and convex behaviour of the graph of the function.

6.23. Determine the Taylor series centered at the origin of function

(a) $y = \frac{1}{(1+x)^2}, \quad x \in (-1, 1)\,;$
(b) $y = \operatorname{arctg} x, \quad x \in (-1, 1).$

Solution. Case (a). We'll use the formula

$$\tfrac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n = \sum_{n=0}^{\infty} (-1)^n x^n, \quad x \in (-1, 1)$$

abou the sum of a geometric series. By differentiating it, we obtain

$$-\tfrac{1}{(1+x)^2} = \left(\sum_{n=0}^{\infty} (-1)^n x^n\right)' = \sum_{n=1}^{\infty} (-1)^n\, n\, x^{n-1}, \quad x \in (-1, 1)\,,$$

with $\left(x^0\right)' = 0$, thus the lower index is $n = 1$. We can see that

$$\tfrac{1}{(1+x)^2} = \sum_{n=1}^{\infty} (-1)^{n+1}\, n\, x^{n-1}, \quad x \in (-1, 1)\,.$$

Case (b). We can express the derivative of function $y = \operatorname{arctg} t$ as

$$(\operatorname{arctg} t)' = \tfrac{1}{1+t^2} = \sum_{n=0}^{\infty} \left(-t^2\right)^n = \sum_{n=0}^{\infty} (-1)^n t^{2n}, \quad t \in (-1, 1)\,.$$

Because for $x \in (-1, 1)$ we have

$$\int_0^x (\operatorname{arctg} t)'\, dt = \operatorname{arctg} x - \operatorname{arctg} 0 = \operatorname{arctg} x$$

and

$$\int_0^x \left(\sum_{n=0}^{\infty} (-1)^n t^{2n}\right) dt = \sum_{n=0}^{\infty} \left((-1)^n \int_0^x t^{2n}\, dt\right) = \sum_{n=0}^{\infty} \tfrac{(-1)^n}{2n+1}\, x^{2n+1}\,,$$

we already have the result

$$\operatorname{arctg} x = \sum_{n=0}^{\infty} \tfrac{(-1)^n}{2n+1}\, x^{2n+1}, \quad x \in (-1, 1)\,.$$

□

6.24. 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

$$\cos t = \sum_{n=0}^{\infty} \tfrac{(-1)^n}{(2n)!}\, t^{2n}, \quad t \in \mathbb{R}$$

implies

$$u \cos u^2 = u \sum_{n=0}^{\infty} \tfrac{(-1)^n}{(2n)!} \left(u^2\right)^{2n} = \sum_{n=0}^{\infty} \tfrac{(-1)^n}{(2n)!}\, u^{4n+1}, \quad u \in \mathbb{R}$$

and then (for $x \in \mathbb{R}$)

$$f(x) = \int_0^x u \cos u^2\, du = \int_0^x \left(\sum_{n=0}^{\infty} \tfrac{(-1)^n}{(2n)!}\, u^{4n+1}\right) du$$
$$= \sum_{n=0}^{\infty} \left(\tfrac{(-1)^n}{(2n)!} \int_0^x u^{4n+1}\, du\right) = \sum_{n=0}^{\infty} \tfrac{(-1)^n}{(2n)!\,(4n+2)}\, x^{4n+2}\,.$$

□

6.25. 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

6.11. Asymptotes of the graph of the function. We'll introduce one more useful utility for sketching the graph of a function. We'll try to figure out the so called *asymptotes*, i.e. the lines, which the values of function f approach. An asymptote at the improper point ∞ is such a line $y = ax + b$, which satisfies

$$\lim_{x\to\infty} (f(x) - ax - b) = 0.$$

We also call it *an asymptote with a slope*. If such an asymptote exists, it satisfies

$$\lim_{x\to\infty} (f(x) - ax) = b$$

and therefore the limit

$$\lim_{x\to\infty} \frac{f(x)}{x} = a$$

exists as well.

Conversely, if the last two limits exist, the limit from the definition of the asymptote exists as well, thus these are sufficient conditions as well.

We can define and compute the asymptote at the improper point $-\infty$ similarly.

This way we can find all the potentional lines satasfying the proporties of asymptotes with a slope. All we have left are potential lines perpendicular the the x axis:

The asyptotes at points $a \in \mathbb{R}$ are lines $x = a$ such that the function f has at least on of the one-sided limits at point a infinite. We also speak of th *asymptotes without a slope*.

For example rational functions have an asymptote in zero points of denominator that aren't zero points of the numerator.

We'll compute at least one simple example: function $f(x) = x + \frac{1}{x}$ has the 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 exactly ± 1 at the improper points, while the limit $f(x) - x = 1/x$ is zero at the improper points.

By differentiating we get

$$f'(x) = 1 - x^{-2}, \quad f''(x) = 2x^{-3}.$$

function $f'(x)$ has two zero points ± 1. At point $x = 1$ the function has a local minimum, at point $x = -1$ a local maximum. The second derivative has no zero points in all its domain $(-\infty, 0) \cup (0, \infty)$, so our function doesn't have any inflection points.

$$\sum_{n=1}^{\infty} n\,(n+1)\,x^n = \sum_{n=1}^{\infty} n\left(x^{n+1}\right)' = \left(\sum_{n=1}^{\infty} n\,x^{n+1}\right)' =$$
$$\left(\sum_{n=1}^{\infty} n\,x^{n-1}\,x^2\right)' = \left[x^2\sum_{n=1}^{\infty}(x^n)'\right]' = \left[x^2\left(\sum_{n=1}^{\infty}x^n\right)'\right]' =$$
$$\left[x^2\left(-1+\sum_{n=0}^{\infty}x^n\right)'\right]' = \left[x^2\left(-1+\tfrac{1}{1-x}\right)'\right]' = \left[x^2\cdot\tfrac{1}{(1-x)^2}\right]' =$$
$$\tfrac{2x}{(1-x)^3}$$

for all $x \in (-1, 1)$. □

6.26. 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.

6.27. Expand the function $\sin^2(x)$ into a power series at point 0 and determine for which real numbers this series converges

6.28. 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. ○

6.29. Expand the function $\ln\sqrt{x}$ into a power series at point 1 and determine for which $x \in \mathbb{R}$ it converges. ○

More problems about Taylor polynomials and series can be found on page 412.

Now we'll state several "classical" problems, in which we'll determine the course of distinct functions.

6.30. Determine the range of function

$$f(x) = \tfrac{e^x-1}{e^x+1}, \quad x \in \mathbb{R}.$$

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

$$\lim_{x\to\infty}\tfrac{e^x-1}{e^x+1} = \lim_{x\to\infty}\tfrac{e^x}{e^x} = 1, \qquad \lim_{x\to-\infty}\tfrac{e^x-1}{e^x+1} = \tfrac{0-1}{0+1} = -1.$$

The inequality

$$f'(x) = \tfrac{2e^x}{(e^x+1)^2} > 0, \quad x \in \mathbb{R}$$

then implies that f is continuous and increasing on $\mathbb{R}$. Hence the range is the interval $(-1, 1)$. □

6.31. State all intervals on which the function $y = e^{-x^2}$, $x \in \mathbb{R}$ is concave. ○

6.32. 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. ○

6.33. Find all asymptotes of function

(a) $y = x\,e^x$;

(b) $y = \frac{(x+3)^3}{(x-2)^3}$

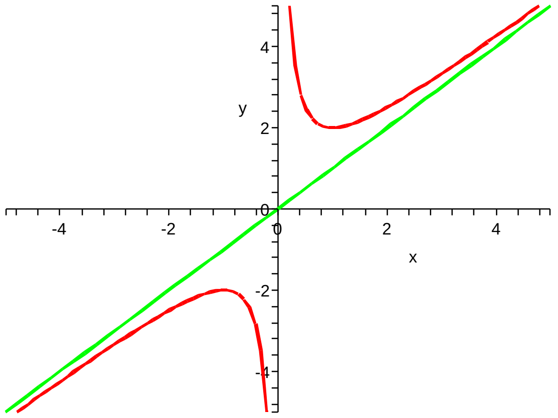
with maximal domain. ○

6.34. State the asymptotes of function

$$y = 2\operatorname{arctg}\left|\tfrac{x}{x^2-1}\right|, \quad x \neq \pm 1\,(x \in \mathbb{R}).$$

○

6.35. Consider function



6.12. Differential of a function. In practical use of differential

calculus, we often work with dependencies between several quantites, say y a x, and the choice of dependant and independant variable is not fixed. The explicit relation $y = f(x)$ with some function f is then only one of several options. Differentiating then expresses, that the immediate change of $y = f(x)$ is proportional to the immediate change of x with a proportion of $f'(x) = \frac{df}{dx}(x)$. This relation is often being written as

$$df(x) = \frac{df}{dx}(x)dx,$$

where we interpret $df(x)$ as a linear map of increments of given $df(x)(\Delta x) = f'(x)\cdot\Delta x$, while $dx(x)(\Delta x) = \Delta x$.

We speak of *the differential of function* f if the approximative property

$$\lim_{\Delta x\to 0}\frac{f(x+\Delta x) - f(x) - df(x)(\Delta x)}{\Delta x} = 0$$

holds. Taylor theorem then implies that a function with bounded derivative f' has a differential df. In particular, that occurs at point x if the first derivative $f'(x)$ exists and is continuous.

If the quantity x is expressed by another quantity t, i.e. $x = g(t)$, and moreover by a function with continuous first derivative again, the the rule for differentiating composite functions tells us the composite function $f \circ g$ has again a differential

$$df(t) = \frac{df}{dx}(x)\frac{dx}{dt}(t)dt.$$

Therefore we can see df as a linear approximation of the given quantity dependant on the increments of the dependant variable, no matter how this dependance is given.

6.13. The curvature of the graph of a function. To train our-

selves in the basic rules for differentiating composite functions etc., we'll dissucss the graph of a smooth function $f(x)$ as a special case of a parametrized curve in a plane for now. We can imagine it as a movement in the plane parametrized by an independant variable x. For an arbitrary point x from the domain of our function, by computing the first derivative we can immediately get the vector $(1, f'(x)) \in \mathbb{R}^2$ that represents the immediate velocity of such a movement. . The tangent line through the point $[x, f(x)]$

$$y = \ln \tfrac{3e^{2x}+e^{x}+10}{e^{x}+1}$$

defined for all real x. Find its asymptotes. ○

6.36. 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

$$(6.4) \qquad \lim_{x\to\pm\infty} \sqrt[3]{|\, x\,|^3 + 1} = \lim_{x\to\pm\infty} \sqrt[3]{|\, x\,|^3} = \lim_{x\to\pm\infty} |\, x\,| = +\infty.$$

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

$$(6.5) \quad f'(x) = \frac{1}{3}\left(x^3+1\right)^{-\frac{2}{3}} 3x^2 = \frac{x^2}{\sqrt[3]{\left(x^3+1\right)^2}} > 0, \qquad x > 0.$$

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 (‖6.4‖)). 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

$$f'(x) = \tfrac{1}{3}\left(-x^3+1\right)^{-\frac{2}{3}}\left(-3x^2\right) = -\tfrac{x^2}{\sqrt[3]{\left(-x^3+1\right)^2}} < 0, \qquad x < 0.$$

When computing $f'(0)$, we can proceed according to the definition or we can use the limits

$$\lim_{x\to 0+} \tfrac{x^2}{\sqrt[3]{(x^3+1)^2}} = 0 = \lim_{x\to 0-} -\tfrac{x^2}{\sqrt[3]{(-x^3+1)^2}}$$

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 (‖6.5‖)

$$f''(x) = \frac{2x\sqrt[3]{(x^3+1)^2} - \frac{2}{3}x^2\sqrt[3]{(x^3+1)^{-1}}(3x^2)}{\sqrt[3]{(x^3+1)^4}},$$

parametrized by this directional vector then represents a linear approximation of the curve.

We've also seen that in the case $f''(x) = 0$ and simultaniously $f'''(x) \neq 0$ the graph of our function intersects the tangent line, which means the tangent line is the best approximation of the curve at the point x up to the second order as well. We usually describe this by saying that the graph of the function f has a *zero curvature* at the point x. While the nonzero values of the first derivative described the speed of the growth (no matter the sign), we can intuitively expect the second derivative will describe the extent of the curvature of the graph. So far we've only seen that the graph of the function is above its tangent for a positive value and below it for a negative one.

We got the tangent at a fixed point $P = [x, f(x)]$ as a limit of the secants, i.e. the lines passing through the points P a $Q = [x + \Delta x, f(x + \Delta x)]$. If we want to approximate the second derivative, we will interpolate the points P and $Q \neq P$ by a circle C_Q, whose center is at the intersection of the perpendicular lines to the tangents at points P and Q. It can be seen from the figure that if the angle between the tangent at a fixed point P and the x axis is α and the angle between the tangent at a fixed point Q and the x axis is $\alpha + \Delta\alpha$, then the angle of the mentioned perpendicular lines will be $\Delta\alpha$ as well. If we denote the radius of our circle by ρ, then the length of the arc between points P and Q will be $\rho\Delta\alpha$. As the point Q approaches a fixed point P, the length of the arc will approach the length s of the studied curve, i.e. the graph of the function $f(x)$, and the circle will approach the circle C_P. Thus we get the basic relation for the limit radius ρ of the circle C_P:

$$\rho = \lim_{\Delta\alpha\to 0} \frac{\Delta s}{\Delta\alpha} = \frac{ds}{d\alpha}.$$

We define the *curvature* of the graph of the function f at a point P as the number $1/\rho$. Zero curvature then corresponds to an infinite radius ρ.

For computing the radius ρ 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.

We can already notice that for an increasing angle θ the length of the arc can either increase or decrease, depending on whether the circle C_Q has the center above or below the graph of the function f. The sign of ρ then reflects whether the function is concave or convex. We also need to think abour the special case when the center "runs off" to infinity in limit, i.e. instead of a circle we get a line again, which is the tangent.

Obviously, we don't have a direct tool to compute the derivative $\frac{ds}{d\alpha}$. However, we know that $\operatorname{tg}\alpha = df/dx$ and by differentiating this equality by x we obtain (using the rule for differentiating composite functions)

$$\frac{1}{(\cos\alpha)^2}\frac{d\alpha}{dx} = f''.$$

On the left hand side we can substitute $\frac{1}{(\cos\alpha)^2} = 1 + (\operatorname{tg}\alpha)^2 = 1 + (f')^2$ which implies (see the rule for differentiating an inverse function)

$$\frac{dx}{d\alpha} = \frac{1 + (\operatorname{tg}\alpha)^2}{f''} = \frac{1 + (f')^2}{f''}.$$

i.e. after a simplification we have

$$(6.6) \qquad f''(x) = \frac{2x}{\sqrt[3]{\left(x^3+1\right)^5}} > 0, \qquad x > 0.$$

Similarly we can compute

$$f''(x) = -\frac{2x\sqrt[3]{\left(-x^3+1\right)^2} - \frac{2}{3}x^2\sqrt[3]{\left(-x^3+1\right)^{-1}}\left(-3x^2\right)}{\sqrt[3]{\left(-x^3+1\right)^4}} =$$

$$-\frac{2x}{\sqrt[3]{\left(-x^3+1\right)^5}} > 0,$$

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 (‖6.6‖), f is strictly convex on the interval $(0, +\infty)$. Also f must be strictly convex on $(-\infty, 0)$. 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 $-\infty$. □

6.37. 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 satisified 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

$$\mathbb{R} \setminus \left\{\tfrac{\pi}{4} + \tfrac{k\pi}{2};\ k \in \mathbb{Z}\right\}.$$

Clearly we have

$$f(-x) = \frac{\cos(-x)}{\cos(-2x)} = \frac{\cos x}{\cos 2x} = f(x)$$

Now we're almost done, because the increment of the length of the arc s in dependance on x is given by the formula

$$\frac{ds}{dx} = (1 + (f')^2)^{1/2}$$

so by using the rule for differentianting a composite function we can compute

$$\rho = \frac{ds}{d\alpha} = \frac{ds}{dx}\frac{dx}{d\alpha} = \frac{(1 + (f')^2)^{3/2}}{f''}.$$

Now we can see the relation between the curvature and the second derivative: the numerator of our fraction is always positive, no matter the value of the first derivative. It's equal to the third power of the size of the tangent vector of the studied curve. The sign of the curvature is therefore given only by the sign of the second derivative, which only confirms our notions about concave and convex points of functions. If the second derivative is zero, the curvature is zero as well. The circle, by which we defined the curvature is called the *osculating circle*.

Try to 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 the easiest, because in these we get the radius of the osculating circle as the reciprocal of the second derivative with the corresponding sign.

6.14. Vector differential calculus. As we've mentioned already in the introduction to chapter five, for our notions about differentiating it was quite essential that we studied functions defined on real numbers and that their values could be added and multiplied by real numbers. That's why we need out functions $f : \mathbb{R} \to V$ to have values in the vector space V. For distinction, we'll call them *vector functions of one real variable* or more briefly *vector functions*.

Now we'll take more interest in real function with values in plane or space, i.e. $f : \mathbb{R} \to \mathbb{R}^2$ and $f : \mathbb{R} \to \mathbb{R}^3$. We talk about (parametrized) curves in plane and space. Similarly, we could work with values in $\mathbb{R}^n$ for any finite dimension n.

For simplification, we'll work in the fixed standard bases e_i in $\mathbb{R}^2$ and $\mathbb{R}^3$, so our curves will be given by doubles or triples of simple real functions of one real variable, respectively. The vector function r in plane or space, respectively, is then 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 a line to the plane or to the space.

In plane it's

$$\frac{dr(t)}{dt}(t) = r'(t) = x'(t)e_1 + y'(t)e_2$$

and similarly in space.

We have to understand the differential of a vector function in this context as well:

$$dr = \left(\frac{dx}{dt}e_1 + \frac{dy}{dt}e_2 + \frac{dz}{dt}e_3\right)dt$$

where we understand the expression on the right hand side in a way the increment of the scalar independant variable t is linearly mapped by multiplying the vector of the derivative and thus obtaining the corresponding increment of the vector quantity r.

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

$$x \in \mathcal{D} := [0, \pi] \smallsetminus \left\{ \tfrac{\pi}{4} + \tfrac{k\pi}{2};\ k \in \mathbb{Z} \right\} = \left[0, \tfrac{\pi}{4}\right) \cup \left(\tfrac{\pi}{4}, \tfrac{3\pi}{4}\right) \cup \left(\tfrac{3\pi}{4}, \pi\right],$$

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

$$\lim_{x \to \frac{\pi}{4}-} \tfrac{\cos x}{\cos 2x} = +\infty, \qquad \lim_{x \to \frac{\pi}{4}+} \tfrac{\cos x}{\cos 2x} = -\infty,$$
$$\lim_{x \to \frac{3\pi}{4}-} \tfrac{\cos x}{\cos 2x} = +\infty, \qquad \lim_{x \to \frac{3\pi}{4}+} \tfrac{\cos x}{\cos 2x} = -\infty.$$

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

$$\begin{aligned} f'(x) &= \frac{-\sin x \cos 2x - 2\cos x\,(-\sin 2x)}{\cos^2 2x} \\ &= \frac{-\sin x\left(\cos^2 x - \sin^2 x\right) + 2\cos x\,(2\sin x \cos x)}{\cos^2 2x} \\ &= \frac{\sin^3 x + 3\cos^2 x \sin x}{\cos^2 2x} = \frac{\left(\sin^2 x + \cos^2 x + 2\cos^2 x\right)\sin x}{\cos^2 2x} \\ &= \frac{\left(2\cos^2 x + 1\right)\sin x}{\cos^2 2x}, \quad x \in \mathcal{D}. \end{aligned}$$

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

$$2\cos^2 x + 1 \ge \cos^2 2x > 0, \quad \sin x > 0, \qquad x \in \mathcal{D} \cap (0, \pi)$$

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 $x \in (-\pi, 0)$, $x \neq -3\pi/4$,

If the $r(t)$ represents a parametrization of a curve, then its derivative is a velocity vector of such defined distance. We studied a special case of the vector $r(t) = te_1 + f(t)e_2$ giving the graph of the function f in the last paragraph. The second derivative then represents the acceleration of such defined movement. Notice that of course the acceleration need not be collinear with the velocity. In fact, in the case of the graph of a function, the acceleration is collinear with the velocity only at the points, where f'' is zero, which corresponds the idea that the acceleration can be collinear only if the curvature of the graph is zero.

6.15. Differentiating composite maps. In linear algebra and geometry there are very useful maps called forms. They have one or more vectors as their arguments and they are linear in each of their aguments. Thus we can define the size of the vectors (the dot product is symmetric bilinear form) or the volume of a parallelepiped (it's a n-linear antisymmetric form, where n is the dimension of the space), see for example the paragraphs 2.44 a 4.22.

Of course, we can use vectors $r(t)$ dependant on a parameter as the arguments of these operations. By a straightforward usage of Leibniz's rule for differentiation of a product of functions we will verify the following

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) Consider differentiable vectors $r_1, \dots, r_k : \mathbb{R} \to \mathbb{R}^n$ and a k–linear form $\Phi : \mathbb{R}^n \times \ldots \times R^n$ on the space $\mathbb{R}^n$. The the derivative of the composed map

$$\varphi(t) = \Phi(r_1(t), \dots, r_k(t))$$

satisfies (generalized Leibniz's) rule

$$\frac{d\varphi}{dt} = \Phi\big(\frac{dr_1}{dt}, r_2, \dots, r_k\big) + \dots + \Phi\big(r_1, \dots, r_{k-1}, \frac{dr_k}{dt}\big).$$

(3) The previous statement remains valid without a change even if Φ also has values in the vector space (and is linear in all its k arguments).

Proof. (1) In linear algebra, it is shown that the linear maps are given by a constant matrix of scalars $A = (a_{ij})$ in a way that

$$\Psi \circ r(t) = \Big(\sum_{i=1}^{n} a_{1i} r_i(t), \dots, \sum_{i=1}^{n} a_{mi} r_i(t)\Big).$$

We now carry out the differentiation by individual coordinates of the result. However, we know, that derivative acts linerly towards scalar linear combinations, see Theorem 5.33. That's why we indeed get the derivative by simply evaluating the original linear map Ψ on the derivative $r'(t)$.

(2) We obtain the second statement analogously. We write out the evaluation of the k–linear form on the vectors $r_1, \dots, r_k$ in the coordinates in this way:

$$\Phi(r_1(t), \dots, r_k(t)) = \sum_{i_1, \dots, i_k = 1}^{n} B_{i_1 \ldots i_k} \cdot (r_1)_{i_1}(t) \ldots (r_k)_{i_k}(t),$$

$x \neq -\pi/4$. 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:

$$f''(x) =$$
$$\frac{\left[4\cos x(-\sin x)\sin x+\left(2\cos^2 x+1\right)\cos x\right]\cos^2 2x-4\cos 2x(-\sin 2x)\left(2\cos^2 x+1\right)\sin x}{\cos^4 2x}$$
$$= \frac{\left[-4\cos x\sin^2 x+2\cos^3 x+\cos x\right]\left(\cos^2 x-\sin^2 x\right)-4(-2\sin x\cos x)\left(2\cos^2 x+1\right)\sin x}{\cos^3 2x}$$
$$=$$
$$\frac{-6\cos^3 x\sin^2 x+2\cos^5 x+\cos^3 x+4\cos x\sin^4 x-\cos x\sin^2 x+16\sin^2 x\cos^3 x+8\sin^2 x\cos x}{\cos^3 2x}$$
$$= \frac{\left[10\sin^2 x\cos^2 x+2\cos^4 x+\cos^2 x+4\sin^4 x+7\sin^2 x\right]\cos x}{\cos^3 2x}, \qquad x \in \mathcal{D}.$$

Note that after a few simplifications, we can also express

$$f''(x) = \frac{\left(3+4\cos^2 x\sin^2 x+8\sin^2 x\right)\cos x}{\cos^3 2x}, \qquad x \in \mathcal{D}$$

or

$$f''(x) = \frac{\left(11-4\cos^4 x-4\cos^2 x\right)\cos x}{\cos^3 2x}, \qquad x \in \mathcal{D}.$$

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

$$f''(x) > 0 \quad \text{pro} \quad x \to 0^+$$

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. □

6.38. 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: $\mathbb{R}^+ \setminus \{1\}$.

where the scalars $B_{i_1 \dots i_k}$ are given as the value of the given form $\Phi(e_{i_1}, \dots, e_{i_k})$ 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's given by finitely many components and we can use the previous notion on each of them □

On the euclidean space $\mathbb{R}^3$, besides the dot product, which assigns a scalar to two vectors, we also have the vector product, which assigns the vector $u \times v \in \mathbb{R}^3$ to vectors u and v, see 4.24. This vector $u \times v$ is orthogonal to both vectors u and v, its size equals the area of the parallelogram determined by the u and v (in this order) and an orientation such that the triple u, v, $u \times v$ is a positively oriented basis.

The previous theorem immediately implies convenient corollaries:

Corollary. *Consider the vectors $u(t)$ and $v(t)$ in the space $\mathbb{R}^3$. The derivatives of their dot product $\langle u(t), v(t)\rangle$ and their vector product $u(t) \times v(t)$ satisfy*

$$\text{(6.1)} \qquad \frac{d}{dt}\langle u(t), v(t)\rangle = \langle u'(t), v(t)\rangle + \langle u(t), v'(t)\rangle$$

$$\text{(6.2)} \qquad \frac{d}{dt}(u(t) \times v(t)) = u'(t) \times v(t) + u(t) \times v'(t)$$

6.16. The curvature of curves. Now we have far more powerful tools for studying curves in amore systematic way than we discussed the curvature of the graphs of functions.

Let's look at the curves $r(t)$ in space and assume they are parametrized in way that their tangent vector always has a unit size, i.e. $\langle r'(t), r'(t)\rangle = 1$ for all t. We say the curve $r(t)$ is *parametrized by the length*. By another differentiation of this unit vector $r'(t)$ we obtain a vector $r''(t)$, for which we'll evaluate (using the symmetry of the dot product)

$$0 = \frac{d}{dt}\langle r'(t), r'(t)\rangle = 2\langle r''(t), r'(t)\rangle$$

and thus the acceleration vector $r''(t)$ is always orthogonal to the velocity vector. That corresponds to the idea that after the choice of a parametrization with a constant size of velocity, the acceleration in the direction of the movement cannot be noticeable, therefore the acceleration must lie in the plane orthogonal to the velocity vector.

If the second derivative is nonzero, we call the normed vector

$$n(t) = \frac{1}{\|r''(t)\|}r''(t)$$

the main normal of the curve $r(t)$. The scalar function $\kappa(t)$ satisfying (at the points where $r''(t) \neq 0$)

$$r''(t) = \kappa(t)n(t)$$

is called *the curvature of the curve* $r(t)$. At the zero points of the second derivative we define $\kappa(t)$ by zero value as well.

At the nonzero points of the curvature the unit vector $b(t) = r'(t) \times n(t)$ is well defined, we call it the *binormal of the curve* $r(t)$. By a direct computation we obtain

$$0 = \frac{d}{dt}\langle b(t), r'(t)\rangle = \langle b'(t), r'(t)\rangle + \langle b(t), r''(t)\rangle$$
$$= \langle b'(t), r'(t)\rangle + \kappa(t)\langle b(t), n(t)\rangle = \langle b'(t), r'(t)\rangle,$$

ii) We'll find the intervals of monotonicity of the function: first we'll find zero points of the derivative:

$$f'(x) = \frac{\ln(x) - 1}{\ln^2(x)} \quad = \quad 0$$

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:

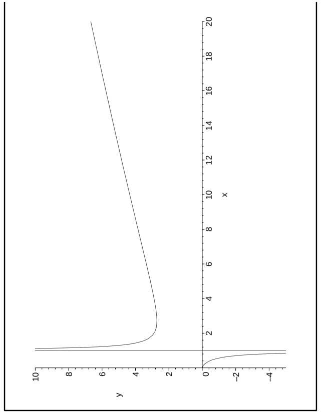
$$k = lim_{x\to\infty} \frac{\frac{x}{\ln(x)}}{x} = \lim_{x\to\infty} \frac{1}{\ln(x)} = 0.$$

If the asymptote exists, its slope must be 0. Let's continue the computation

$$\lim_{x\to\infty} \frac{x}{\ln(x)} - 0 \cdot x = \lim_{x\to\infty} \ln(x) = \infty,$$

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

6.39. Determine the curvature of the ellipse $x^2 + 2y^2 = 2$ at its vertices (‖4.47‖). Also determine the equations of the circles of osculation at these vertices.

which show that the tangent vector to the binormal is orthogonal to both $b(t)$ and $r'(t)$. Therefore it must be a multiple of the vector of the main normal. We write

$$b'(t) = -\tau(t)n(t)$$

and call the scalar function $\tau(t)$ the *torsion of the curve* $r(t)$.

We have not yet computed the speed of change of the main normal, which we can also write as $n(t) = b(t) \times r'(t)$:

$$\begin{aligned} n'(t) &= b'(t) \times r'(t) + \kappa(t)b(t) \times n(t) \\ &= -\tau(t)n(t) \times r'(t) + \kappa(t)(-r'(t)) \\ &= \tau(t)b(t) - \kappa(t)r'(t). \end{aligned}$$

Successively, for all points with nonzero second derivative of the curve $r(t)$ parametrized by the length of the arc, we derived an importnat basis $(r'(t), n(t), b(t))$, called *the Frenet frame* in the classical literature and simultaneously in this base we expressed the derivatives of its components by the form of the so called *Frenet–Serret formulas*

$$\frac{dr'}{dt}(t) = \kappa(t)n(t), \quad \frac{dn}{dt}(t) = \tau(t)b(t) - \kappa(t)r'(t)$$
$$\frac{db}{dt}(t) = -\tau(t)n(t).$$

Notice that if the curve $r(t)$ still lies in one plane, then its torsion is an identically zero function. In fact, the converse is true as well. We won't prove it here, but it is a corollary of a classical result of geometric theory of curves:

Two curves in a space parametrized by the length of the 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 funcitons κ a τ there exists a smooth curve with these parameters.

We won't prove this result here, the persons concerned can find the thorough version in[**?**].

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, by computing the derivative of the composite function using the differential of the length of the arc for the graph of a function of form

$$dt = (1 + (f_x)^2)^{1/2}dx, \qquad dx = (1 + (f_x)^2)^{-1/2}dt$$

(here we write $f_x = \frac{df}{dx}$) we obtain this relation for our unir tangent vector of the graph of a cruve

$$r'(t) = (x'(t), y'(t)) = \big((1 + (f_x)^2)^{-1/2}, f_x(1 + (f_x)^2)^{-1/2}\big)$$

and by fairly not well arranged, but similar computation of the second derivative and its size we indeed obtain

$$\kappa^2 = \|r''\|^2 = (\tfrac{d^2 f}{dx^2})^2(1 + (f_x)^2)^{-3}.$$

6.17. The approximations of derivatives and the asymptotic estimations.

In the beggining of this textbook in paragraphs 1.3, 1.9 and further we discussed how to express the value of a function by changes, i.e. differentions. In the next part of the text we will similaely reconstruct the function f using its derivatives, i.e. immediate changes. Before that though, let's stop at the connections between

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 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 6.13, 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 6.13 and 6.16 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 6.13. The normal line is heading to $-\infty$ along the x axis at point $[\sqrt{2}, 0]$, thus the center of the osculation circle will be on the x axis on the other side at distance $\frac{\sqrt{2}}{2}$, hence at the point $[\sqrt{2} - \frac{\sqrt{2}}{2}, 0] = [\frac{\sqrt{2}}{2}, 0]$. In total, the equation of the circle of osculation at vertice $[\sqrt{2}, 0]$ will be $(x - \frac{\sqrt{2}}{2})^2 + y^2 = \frac{1}{2}$. The curvature at both of these vertices equals $\sqrt{2}$.

derivatives and differentions. The key to this will be the Taylor series with a remainder.

Suppose that for some (sufficiently) differentiable function $f(x)$ defined on the interval $[a, b]$, we know the values $f_i = f(x_i)$ at the points $x_0 = a, x_1, x_2, \dots, x_n = b$, while all indices $i = 1, \dots, n$ satisfy $x_i - x_{i-1} = h > 0$ for some constant h. Write the Taylor expansion of function f in the form

$$f(x_i \pm h) = f_i \pm hf'(x_i) + \frac{h^2}{2}f''(x_i) \pm \frac{h^3}{3!}f^{(3)}(x_i) + \dots$$

We know that if we finish the expansion by a term of order k in h, i.e. an expression containing h^k, then the actual error will be bounded by

$$\frac{h^{k+1}}{(k+1)!}f^{(k+1)}(x)$$

on the interval $[x_i - h, x_i + h]$. If the $(k + 1)$–the derivative f is continuous, we can approximate it by a constant. Then we can see that 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*.

Definition. We say the expression $G(h)$ is asymptoticall equal to $F(h)$ for $h \to 0$ and write $G(h) = O(F(h))$, if the finite limit

$$\lim_{h \to 0} \frac{G(h)}{F(h)} = a \in \mathbb{R}$$

exists.

Denote the sought estimations of the values of the derivatives of $f(x)$ at the points x_i as $f_i^{(j)}$ and write the Taylor expansion briefly in this way:

$$f_{i\pm1} = f_i \pm f_i'h + \frac{f_i''}{2}h^2 \pm \frac{f_i'''}{6}h^3 + \dots$$

For the approximations of the first derivative we can immediately use three different differences computed from the Taylor expansion:

$$f_i^{(1)} = \frac{f_{i+1} - f_{i-1}}{2h} - \frac{h^2}{3!}f^{(3)}(x_i) - \dots$$

$$f_i^{(1)} = \frac{f_{i+1} - f_i}{h} - \frac{h}{2!}f''(x_i) + \dots$$

$$f_i^{(1)} = \frac{f_i - f_{i-1}}{h} + \frac{h}{2!}f''(x_i) + \dots$$

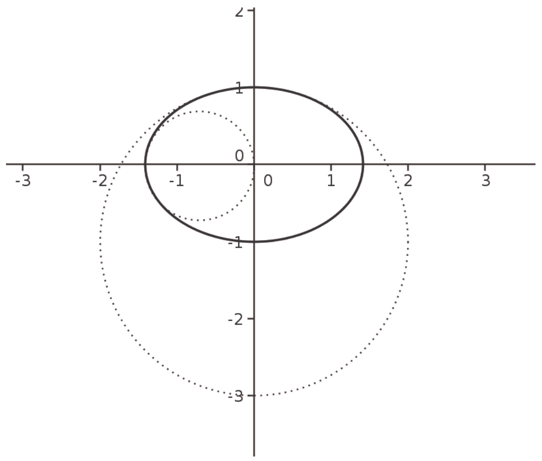
when we only substracted the respective polynomials. Then we obtain a numerical representation of the first derivative. The first of them has an asymptotic estimation of the error of

$$f^{(1)} = \frac{f_{i+1} - f_{i-1}}{2h} + O(h^2),$$

the other two have $O(h)$. We call them *the central difference*, *the forward difference* and *the backward difference*. Surprisingly, the central difference one digit place better than the other two.

We can proceed the same way when approximating the second derivative. To be able to compute $f''(x_i)$ from a suitable combination of the Taylor polynomials, we need to cancel both the first derivatives and the value at x_i. The simpliest combination cancels all the odd derivatives as well:

$$f_i^{(2)} = \frac{f_{i+1} - 2f_i + f_{i+1}}{h^2} + \frac{h^2}{12}f^{(4)}(x_i) + \dots.$$



□

6.40. 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

First, several easy examples that everyone should handle.

6.41. 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

(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 $\int \frac{f'(x)}{f(x)}\,dx = \ln|f(x)| + C$.

□

6.42. Compute the indefinite integral

$$\int \left(7^x + 4\,e^{\frac{2x}{3}} - \frac{1}{2^x} + 9\sin 5x + 2\cos\frac{x}{2} - \frac{3}{\cos^2 x} + \frac{1}{3-x}\right)dx$$

pro $x \neq 3$, $x \neq \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$.

Solution. Only by combining the earlier derived formulas, we obtain

We call this *the differention of the second order* and just like the central first differention, the asymptotic estimation is one digit place better than we would expect at first glance:

$$f_i^{(2)} = \frac{f_{i+1} - 2f_i + f_{i+1}}{h^2} + O(h^2).$$

2. Integration

6.18. Newton integral. Now we'll take interest in the reverse procedure than we did for differentiating. We'll want to reconstruct the actual values of some function using its immediate changes. If we consider the given function $f(x)$ the derivative of an unknown function $F(x)$ then at the differential level we can write

$$dF = f(x)dx.$$

We call the function F *the primitive function* or *the indefinite integral* of the function f and 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 $[a, b]$ except for an additive constant.*

Proof. The statement follows immediately from Lagrange's mean value theorem, see 5.38. Indeed, if $F'(x) = G'(x) = f(x)$ on the whole interval $[a, b]$, the function $(F - G)(x)$ has a zero derivative in all points c of the interval $[a, b]$. Then the mean value theorem implies that for all points x in this interval,

$$F(x) - G(x) = F(a) - G(a) + 0 \cdot (x - a).$$

Then the difference of the values of functions F a G must be the same on the interval $[a, b]$. □

The previous lemma leads us to this notation for the indefinite integral:

$$F(x) = \int f(x)dx + C$$

with an unknown constant C.

We can also consider the value of real function $f(x)$ as an immediate increment of the area bounded by the graph of the function f and the x axis and try to find the size of this area between boundary values a a b of some interval. Let's try to connect this concept with the indefinite integral. Suppose we know a real function and its indefinite integral $F(x)$, i.e. $F'(x) = f(x)$ on the interval $[a, b]$.

If we divide the interval $[a, b]$ to n parts by choosing the points

$$a = x_0 < x_1 < \cdots < x_n = b$$

and approximate the values of the derivatives at the points x_i by the expressions

$$f(x_i) = F'(x_i) \simeq \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i},$$

by summing over all the intervals of our partition, we obtain an estimation of the sought size of the area:

$$\int \left(7^x + 4\,\mathrm{e}^{\frac{2x}{3}} - \tfrac{1}{2^x} + 9\sin 5x + 2\cos\tfrac{x}{2} - \tfrac{3}{\cos^2 x} + \tfrac{1}{3-x}\right) dx =$$
$$\tfrac{7^x}{\ln 7} + 6\,\mathrm{e}^{\frac{2x}{3}} + \tfrac{1}{2^x \ln 2} - \tfrac{9}{5}\cos 5x + 4\sin\tfrac{x}{2} - 3\operatorname{tg} x - \ln|\,3 - x\,| + C.$$

□

For expressing the following integrals, we'll use the method of integration by parts (see 6.20).

6.43. Compute $\int x\cos x\ \mathrm{d}x,\ x \in \mathbb{R}$ and $\int \ln x\ dx,\ x > 0$;

Solution.

$$\begin{aligned}\int \ln x\,\mathrm{d}x &= \begin{vmatrix} u = \ln x & u' = \frac{1}{x} \\ v' = 1 & v = x \end{vmatrix} \\ &= x\ln x - \int 1\,\mathrm{d}x = x\ln x - x + C.\end{aligned}$$

$$\begin{aligned}\int x\cos x\ \mathrm{d}x &= \begin{vmatrix} u = x & u' = 1 \\ v' = \cos x & v = \sin x \end{vmatrix} = x\sin x - \int \sin x\,\mathrm{d}x \\ &= x\sin x + \cos x + C.\end{aligned}$$

□

6.44. Using integration by parts, compute

(a) $\int \left(x^2+1\right)\mathrm{e}^{-x}\,dx,\ x \in \mathbb{R}$,
(b) $\int (2x-1)\ln x\ dx,\ x > 0$,
(c) $\int \operatorname{arctg} x\ dx,\ x \in \mathbb{R}$,
(d) $\int \mathrm{e}^x \sin x\ dx,\ x \in \mathbb{R}$,

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

$$a \in (0,1)\cup(1,+\infty), \quad b, c \in \mathbb{R}\smallsetminus\{0\}, \quad n \in \mathbb{N}, \quad k > 0.$$

Thus we know that

(a)

$$\begin{gathered}\int \left(x^2+1\right)\mathrm{e}^{-x}\,dx = \left|\begin{array}{c|c} F(x) = x^2+1 & F'(x) = 2x \\ G'(x) = \mathrm{e}^{-x} & G(x) = -\mathrm{e}^{-x} \end{array}\right| = \\ -\left(x^2+1\right)\mathrm{e}^{-x} + \int 2x\,\mathrm{e}^{-x}\,dx = \\ \left|\begin{array}{c|c} F(x) = 2x & F'(x) = 2 \\ G'(x) = \mathrm{e}^{-x} & G(x) = -\mathrm{e}^{-x} \end{array}\right| = \\ -\left(x^2+1\right)\mathrm{e}^{-x} - 2x\,\mathrm{e}^{-x} + \int 2\,\mathrm{e}^{-x}\,dx = -\left(x^2+1\right)\mathrm{e}^{-x} - \\ 2x\,\mathrm{e}^{-x} - 2\,\mathrm{e}^{-x} + C = -\mathrm{e}^{-x}\left(x^2+2x+3\right) + C;\end{gathered}$$

(b)

$$\begin{gathered}\int (2x-1)\ln x\ dx = \left|\begin{array}{c|c} F(x) = \ln x & F'(x) = 1/x \\ G'(x) = 2x-1 & G(x) = x^2-x \end{array}\right| = \\ \left(x^2-x\right)\ln x - \int \tfrac{x^2-x}{x}\,dx = \left(x^2-x\right)\ln x + \int 1 - x\,dx = \\ \left(x^2-x\right)\ln x + x - \tfrac{x^2}{2} + C;\end{gathered}$$

(c)

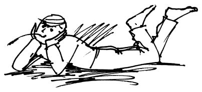
$$\begin{aligned}\sum_{i=0}^{n-1} f(x_i)\cdot(x_{i+1}-x_i) &\simeq \sum_{i=0}^{n-1} \frac{F(x_{i+1}) - F(x_i)}{x_{i+1}-x_i}\cdot(x_{i+1}-x_i) \\ &= F(b) - F(a).\end{aligned}$$

Therefore we can expect that for "nice enough" functions $f(x)$, the size of the area bounded by the graph of the function and the x axis can truly be calculated as a difference of the values of the primitive function at the boundary points of the interval. This procedure is called *Newton integral*. We write

$$\int_a^b f(x)dx = [F(x)]_a^b = F(b) - F(a)$$

and also speak of the (Newton) *definite integral* within the bounds a, b.

In the case of a complex function f, the real and the imaginary part of its indefinite integral is uniquely determined by the real and the imaginary part of f, so with no further comments we'll only work with real functions from now on and we'll come back to complex ones in applications as needed.



6.19. Integration "by heart". Before we'll clarify how the Newton integral is connected to the size of an area a eventually how to use it for simulations of practical problems, we'll show several procedures of computing the Newton integral. We'll only use our knowledge of differentiation.

The most easy case is the one when we can see the derivative in the integrated function flat out. To do that in the simple cases, it suffices to read the tables for function derivatives in our menagerie from the other side. This way we get f.ex. the following statements for all $a \in \mathbb{R}$ and $n \in \mathbb{Z}, n \neq -1$:

$$\int a\,dx = ax + C$$

$$\int ax^n\ dx = \tfrac{a}{n+1}x^{n+1} + C$$

$$\int \mathrm{e}^{ax}\ dx = \tfrac{1}{a}\,\mathrm{e}^{ax} + C$$

$$\int \frac{a}{x}\,dx = a\ln x + C$$

$$\int a\cos(bx)\,dx = \tfrac{a}{b}\sin(bx) + C$$

$$\int a\sin(bx)\,dx = -\tfrac{a}{b}\cos(bx) + C$$

$$\int a\cos(bx)\sin^n(bx)\,dx = \tfrac{a}{b(n+1)}\sin^{n+1}(bx) + C$$

$$\int a\sin(bx)\cos^n(bx)\,dx = -\tfrac{a}{b(n+1)}\cos^{n+1}(bx) + C$$

$$\int a\operatorname{tg}(bx)\,dx = -\frac{a}{b}\ln(\cos(bx)) + C$$

$$\int \frac{a}{a^2+x^2}\,dx = \operatorname{arctg}\left(\tfrac{x}{a}\right) + C$$

$$\int \frac{-1}{\sqrt{a^2-x^2}}\,dx = \arccos\left(\tfrac{x}{a}\right) + C$$

$$\int \frac{1}{\sqrt{a^2-x^2}}\,dx = \arcsin\left(t\frac{x}{a}\right) + C.$$

$$\int \operatorname{arctg} x\, dx = \left| \begin{array}{l|l} F(x) = \operatorname{arctg} x & F'(x) = \frac{1}{1+x^2} \\ G'(x) = 1 & G(x) = x \end{array} \right| = \\ x \operatorname{arctg} x - \int \frac{x}{1+x^2}\, dx = x \operatorname{arctg} x - \frac{1}{2} \int \frac{2x}{1+x^2}\, dx = \\ x \operatorname{arctg} x - \frac{1}{2} \ln\left(1+x^2\right) + C;$$

(d)

$$\int e^x \sin x\, dx = \left| \begin{array}{l|l} F(x) = e^x & F'(x) = e^x \\ G'(x) = \sin x & G(x) = -\cos x \end{array} \right| = \\ -e^x \cos x + \int e^x \cos x\, dx = \\ \left| \begin{array}{l|l} F(x) = e^x & F'(x) = e^x \\ G'(x) = \cos x & G(x) = \sin x \end{array} \right| = \\ -e^x \cos x + e^x \sin x - \int e^x \sin x\, dx,$$

which implies

$$\int e^x \sin x\, dx = \frac{1}{2} e^x (\sin x - \cos x) + C.$$

□

For expressing the following integrals, it's convenient to use the substitution method (see 6.21).

6.45. 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)

$$\int \sqrt{2x-5}\, dx = \left| \begin{array}{c} t = 2x-5 \\ dt = 2\, dx \end{array} \right| = \frac{1}{2} \int \sqrt{t}\, dt = \frac{1}{3} t^{\frac{3}{2}} + C = \\ \frac{1}{3}\sqrt{(2x-5)^3} + C;$$

(b)

$$\int \frac{(7+\ln x)^7}{x}\, dx = \left| \begin{array}{c} t = 7+\ln x \\ dt = \frac{1}{x}\, dx \end{array} \right| = \int t^7\, dt = \frac{t^8}{8} + C = \\ \frac{(7+\ln x)^8}{8} + C;$$

(c)

$$\int \frac{\cos x}{(1+\sin x)^2}\, dx = \left| \begin{array}{c} t = 1+\sin x \\ dt = \cos x\, dx \end{array} \right| = \int \frac{dt}{t^2} = -\frac{1}{t} + C = \\ -\frac{1}{1+\sin x} + C;$$

(d)

$$\int \frac{\cos x}{\sqrt{1+\sin^2 x}}\, dx = \left| \begin{array}{c} t = \sin x \\ dt = \cos x\, dx \end{array} \right| = \int \frac{1}{\sqrt{1+t^2}}\, dt = \\ \left| \begin{array}{c} u = t + \sqrt{1+t^2} > 0 \\ du = \left(1 + \frac{t}{\sqrt{1+t^2}}\right) dt \\ \frac{du}{t+\sqrt{1+t^2}} = \frac{1}{\sqrt{1+t^2}}\, dt \end{array} \right| = \int \frac{1}{u}\, du = \ln u + C = \\ \ln\left(t + \sqrt{1+t^2}\right) + C = \ln\left(\sin x + \sqrt{1+\sin^2 x}\right) + C.$$

□

In all the cases it's necessary to think through the domain on which the indefinite integral is well defined.

We can relatively easily add another rules by simple observations of suitable structure of the integrated functions to these table rules. F.ex.

$$\int \frac{f'(x)}{f(x)}\, dx = \ln|f(x)| + C$$

for all continuously differentiable functions f on the intervals where they are nonzero. Of course, the rules for differentianting a sum of differentiable functions and constant multiples of differentiable functions imply that analogous rules hold for the indefinite integral.

6.20. Integration by parts. The computation of the integral using a primitive function (the indefinite integral), along with rule

$$(F \cdot G)'(t) = F'(t) \cdot G(t) + F(t) \cdot G'(t)$$

for differentiating a product of functions, gives us the following formula for the indefinite integral

$$F(x) \cdot G(x) + C = \int F'(x)G(x)\, dx + \int F(x)G'(x)\, dx.$$

This formula is usually used in a way that one of the integrals on the right hand side is the one we want to compute, while we can compute the other one more easily.

The principle is best shown on an example. Let's compute

$$I = \int x \sin x\, dx.$$

In this case a choice $F(x) = x$, $G'(x) = \sin x$ will help. Then $G(x) = -\cos x$ and therefore

$$I = -x \cos x - \int -\cos x\, dx = -x \cos x + \sin x + C.$$

A common trick is also using this procedure for $F'(x) = 1$:

$$\int \ln x\, dx = \int 1 \cdot \ln x\, dx = x \ln x - \int \frac{1}{x} x\, dx = x \ln x - x + C.$$

6.21. Integration by substitution. Another useful procedure is derived from differentiating composite functions. If

$$F'(y) = f(y), \qquad y = \varphi(x),$$

for a differentiable function φ 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 then get the originally desired primitive function. More often, we write this fact in this way:

$$\int f(y)\, dy = \int f(\varphi(x))\varphi'(x)\, dx$$

and we talk of substituting the variable y. 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

$$dy = \varphi'(x)\, dx,$$

6.46. Determine the integrals

a) $\int \frac{\mathrm{d}x}{\sin^2(x)-\cos^2(x)}$,

b) $\int x^2\sqrt{2x+1}\,\mathrm{d}x$.

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,

$$\int x^2\sqrt{2x+1}\,\mathrm{d}x =$$

$$= \frac{1}{7}x^2(2x+1)^{\frac{3}{2}} - \frac{2}{7}\int x\sqrt{2x+1} =$$

$$\left| \begin{array}{ll} u = x & u' = 1 \\ v' = \sqrt{2x+1} & v = \frac{1}{3}\sqrt{2x+1} \end{array} \right|$$

$$= \frac{1}{7}x^2(2x+1)^{\frac{3}{2}} - \frac{2}{7}\left(\frac{1}{3}x\sqrt{2x+1} - \frac{1}{3}\int (2x+1)^{\frac{3}{2}}\,\mathrm{d}x\right) =$$

$$= \frac{1}{7}x^2(2x+1)^{\frac{3}{2}} - \frac{2}{21}x\sqrt{2x+1} + \frac{2}{105}(2x+1)^{\frac{5}{2}} =$$

$$= \frac{1}{7}x^2(2x+1)^{\frac{3}{2}} - \frac{2}{35}x(2x+1)^{\frac{3}{2}} + \frac{2}{105}(2x+1)^{\frac{3}{2}} + C$$

□

6.47. 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) $\int 6\sin 5x + \cos\frac{x}{2} + 2\,\mathrm{e}^{\frac{2x}{3}}\,dx,\ x \in \mathbb{R}.$

Solution. Case (a). We can immediately determine

$$\int \frac{1}{\sqrt[3]{x}}\,dx = \int x^{-1/3}\ dx = \frac{x^{\frac{2}{3}}}{\frac{2}{3}} + C = \frac{3}{2}\sqrt[3]{x^2} + C,$$

which corresponds the relation between the integrated quantities

$$f(y)dy = f(\varphi(x))\varphi'(x)dx.$$

As an example, we'll verify the penultimate integral in the list in 6.20 using this method. For the integral

$$I = \int \frac{1}{\sqrt{1-x^2}}\,dx$$

we'll choose the substitution $x = \sin t$. Then $dx = \cos t\,dt$ and we obtain

$$I = \int \frac{1}{\sqrt{1-\sin^2 t}}\cos t\,dt = \int \frac{1}{\sqrt{\cos^2 t}}\cos t\,dt$$

$$= \int dt = t + C.$$

By reverse substitution $t = \arcsin x$ we get the already known relation $I = \arcsin x + C$.

While substituting, we need to be aware of the actual existence of the inverse function to $y = \varphi(x)$; while evaluating a definite Newton integral we also need to correctly recalculate the bounds of integration. The problems with the domains of the inverse functions can sometimes be avoided by dividing the integraiton into several intervals.

6.22. Integration by reduction to reccurences. Often the use of subtitutions and integrating by parts leads to reccurent relations, from which we can evaluate the desired integrals. We'll illustrate this on an examples. By integration by parts, we 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 $\sin^2 x = 1 - \cos^2 x$, we get

$$mI_m = \cos^{m-1} x\sin x + (m-1)I_{m-2}$$

and the initial values are

$$I_0 = x, \quad I_1 = \sin x.$$

The integrals in which the integrated function depends on expressions of the form (x^2+1) can be often reduced to these types of integrals using the substitution $x = \operatorname{tg} t$. Indeed, f.ex. for

$$J_k = \int \frac{dx}{(x^2+1)^k}$$

the mentioned substitution gets us (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

$$J_2 = \frac{1}{2}(\cos t\sin t + t) = \frac{1}{2}\left(\frac{\operatorname{tg} t}{1+\operatorname{tg}^2 t} + t\right)$$

and therefore after the reverse substitution $t = \operatorname{arctg} x$

$$J_2 = \frac{1}{2}\left(\frac{x}{1+x^2} + \operatorname{arctg} x\right) + C.$$

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. 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 $a \in \mathbb{R} \setminus \{0\}$ 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

$$\int \operatorname{tg}^2 x\, dx = \int \tfrac{\sin^2 x}{\cos^2 x}\, dx = \int \tfrac{1-\cos^2 x}{\cos^2 x}\, dx =$$
$$\int \tfrac{1}{\cos^2 x}\, dx - \int 1\, dx = \operatorname{tg} x - x + C,$$

where we helped ourselves by the knowledge of the derivative

$$(\operatorname{tg} x)' = \tfrac{1}{\cos^2 x}, \qquad x \neq \tfrac{\pi}{2} + k\pi,\ k \in \mathbb{Z}.$$

Case (c). It suffices to realize that this is a special case of the formula

$$\int \tfrac{f'(x)}{f(x)}\, dx = \ln |\, f(x)\,| + C,$$

which can be verified directly by differentiation

$$(\ln |\, f(x)\,| + C)\,' = \big(\ln\left[\pm f(x)\right]\big)' + (C)' = \tfrac{[\pm f(x)]'}{\pm f(x)} = \tfrac{\pm f'(x)}{\pm f(x)} = \tfrac{f'(x)}{f(x)}.$$

Hence

$$\int \tfrac{\cos x}{1+\sin x}\, dx = \ln\ (1 + \sin x) + C.$$

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

$$\int 6\sin 5x + \cos\tfrac{x}{2} + 2\,\mathrm{e}^{\frac{2x}{3}}\, dx = -\tfrac{6}{5}\cos 5x + 2\sin\tfrac{x}{2} + 3\,\mathrm{e}^{\frac{2x}{3}} + C.$$

□

6.48. Determine

(a) $\int \frac{x}{\cos^2 x}\, dx,\ x \neq \frac{\pi}{2} + k\pi,\ k \in \mathbb{Z};$

(b) $\int x^2\, \mathrm{e}^{-3x}\ dx,\ x \in \mathbb{R};$

(c) $\int \cos^2 x\ dx,\ x \in \mathbb{R}.$

Solution. Case (a). Using integration by parts, we obtain

$$\int \tfrac{x}{\cos^2 x}\, dx = \left|\begin{array}{l|l} F(x) = x & F'(x) = 1 \\ G'(x) = \frac{1}{\cos^2 x} & G(x) = \operatorname{tg} x \end{array}\right| = x \operatorname{tg} x - \int \operatorname{tg} x\, dx =$$
$$x \operatorname{tg} x + \int \tfrac{-\sin x}{\cos x}\, dx = x \operatorname{tg} x + \ln|\cos x\,| + C.$$

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

While evaluating definite integrals, we can compute the whole reccurence straight out after avaluating in the given bounds. F.ex. it can be seen immediately that while doing integration over the interval $[0, 2\pi]$, our 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 $m = 2n$ we obtain the resulr

$$\int_0^{2\pi} \cos^{2n} x\, dx = \frac{(2n-1)(2n-3)\dots 3\cdot 1}{2n(2n-2)\dots 2} 2\pi,$$

outright, while for odd m it's always zero (as could be guessed from the graph of the function $\cos x$).

6.23. Integration of rational functions. While doing integration of rational functions, we can use several simplifications. Particularly in the case the degree of the polynomial f in the numerator is greater or equal to the degree of the polynomial g in the denominator, it's sensible to carry out the division with a remainder outright (see the paragraph 5.2) and reduce the integration to a sum of two integrals. The first one will be an integration of a polynomial and the second one an integration of an expression f/g with degree of g strictly greater than the degree of f (such functions are called *proper rational functions*).

Indeed, we can achieve this by simple division of the polynomial:

$$f = q \cdot g + h, \quad \frac{f}{g} = q + \frac{h}{g}.$$

Thus we can assume without loss of generality that degree of g is strictly greater than the degree of f. We'll show another procedure on a simple example. Let's try to analyse how to get the result

$$\frac{f(x)}{g(x)} = \frac{4x+2}{x^2+3x+2} = \frac{-2}{x+1} + \frac{6}{x+2},$$

which we can integrate directly:

$$\int \frac{4x+2}{x^2+3x+2}\, dx = -2\ln|x+1| + 6\ln|x+2| + C.$$

First off, by modifying the sum of the fractions to a common denominator we can verify this equality easily. Conversely, if we know our expression can be written in the form

$$\frac{4x+2}{x^2+3x+2} = \frac{A}{x+1} + \frac{B}{x+2},$$

we only need to compute the cofficients A and B. We can obtain equations for them by multiplying both side by the polynomial $x^2 + 3x + 2$ from the denominator and comparing coefficients of the individual powers of x in the resulting polynomials on both sides:

$$4x + 2 = A(x+2) + B(x+1) \implies 2A + B = 2,\ A + B = 4.$$

This is where our decomposition comes from. It's called *decomposition into partial fractions*.

This elementary procedure can easily be generalized. It's a purely algebraic notion based upon properties of polynomials, which we'll come back to in chapter **??**.

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

$$\int x^2\, e^{-3x}\, dx = \left|\begin{array}{l|l} F(x) = x^2 & F'(x) = 2x \\ G'(x) = e^{-3x} & G(x) = -\frac{1}{3}\, e^{-3x} \end{array}\right| = -\tfrac{1}{3}\, x^2\, e^{-3x} + \tfrac{2}{3} \int x\, e^{-3x}\, dx$$

and furthermore

$$\int x\, e^{-3x}\, dx = \left|\begin{array}{l|l} F(x) = x & F'(x) = 1 \\ G'(x) = e^{-3x} & G(x) = -\frac{1}{3}\, e^{-3x} \end{array}\right| = -\tfrac{1}{3}\, x\, e^{-3x} + \tfrac{1}{3} \int e^{-3x}\, dx = -\tfrac{1}{3}\, x\, e^{-3x} - \tfrac{1}{9}\, e^{-3x} + C.$$

In total, we have

$$\int x^2\, e^{-3x}\, dx = -\tfrac{1}{3}\, x^2\, e^{-3x} - \tfrac{2}{9}\, x\, e^{-3x} - \tfrac{2}{27}\, e^{-3x} + C = -\tfrac{1}{3} e^{-3x} \left(x^2 + \tfrac{2}{3}x + \tfrac{2}{9}\right) + C.$$

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

$$\int \cos^2 x\, dx = \int \cos x \cdot \cos x\, dx =$$
$$\left|\begin{array}{l|l} F(x) = \cos x & F'(x) = -\sin x \\ G'(x) = \cos x & G(x) = \sin x \end{array}\right| = \cos x \cdot \sin x + \int \sin^2 x\, dx =$$
$$\cos x \cdot \sin x + \int 1 - \cos^2 x\, dx =$$
$$\cos x \cdot \sin x + \int 1\, dx - \int \cos^2 x\, dx = \cos x \cdot \sin x + x - \int \cos^2 x\, dx.$$

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.

$$\int \cos^2 x\, dx = \frac{1}{2}\, (x + \sin x \cdot \cos x) + C. \tag{6.7}$$

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.

We emphasise that usually suitable simplifications or substitutions lead to the result faster than integration by parts. For example, by using the identity

$$\cos^2 x = \tfrac{1}{2}\, (1 + \cos 2x)\,, \qquad x \in \mathbb{R}$$

we easily obtain

$$\int \cos^2 x\, dx = \int \tfrac{1}{2}\, dx + \int \tfrac{1}{2} \cos 2x\, dx = \tfrac{x}{2} + \tfrac{\sin 2x}{4} + C = \tfrac{x}{2} + \tfrac{2 \sin x \cos x}{4} + C = \tfrac{1}{2}\, (x + \sin x \cdot \cos x) + C.$$

□

6.49. 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) $\frac{1-\sqrt[3]{x}+\sqrt{x}}{\sqrt[6]{x^5}+x}\ dx,\ x > 0.$

Suppose the denominator $g(x)$ and the numerator $f(x)$ don't share any real or complex roots and that $g(x)$ has exactly n distinct real roots $a_1, \dots, a_n$. Then the points $a_1, \dots a_n$ are exactly all the discontinuities of the function $f(x)/g(x)$.

For simplifying the notion first write $g(x)$ as the product

$$g(x) = p(x)q(x)$$

of two coprime polynomials. By Bezout identity (see **??**), which is a corollary of ordinary polynomial division with a remainder, there exist polynomials $a(x)$ and $b(x)$ of degrees strictly lower than the degree of g such that

$$a(x)p(x) + b(x)q(x) = 1.$$

Multiplying this equality by the quotient $f(x)/g(x)$, we obtain

$$\frac{f(x)}{g(x)} = \frac{a(x)}{q(x)} + \frac{b(x)}{p(x)}.$$

Now suppose our polynomial $g(x)$ has only real roots, therefore it has a unique factorization $(x - a_i)^{n_i}$, where n_i are the multiplicities of the roots a_i, $i = 1, \dots, k$. By a sequential use of the previous procedure with coprime polynomials $p(x)$ and $q(x)$, we get a representation of $f(x)/g(x)$ as a sum of fractions of the form

$$\frac{r_1(x)}{(x - a_1)^{n_1}} + \cdots + \frac{r_k(x)}{(x - a_k)^{m_j}},$$

where the degrees of the polynomials $r_i(x)$ are strictly lesser than the degrees of the denominators. Each of them can be very easily represented as a sum

$$\frac{r(x)}{(x - a)^n} = \frac{A_1}{x - a} + \frac{A_2}{(x - a)^2} + \cdots + \frac{A_n}{(x - a)^n},$$

if we start from the highest powers of the polynomial $r(x)$ and sequentially compute $A_1, A_2, \dots$ by suitable adding and removing of summands in the numerator. F.ex.

$$\frac{5x - 16}{(x - 2)^2} = 5\frac{x - 2}{(x - 2)^2} - 6\frac{1}{(x - 2)^2} = \frac{5}{x - 2} + \frac{6}{(x - 2)^2}.$$

Now we need to handle the case, where there are not enough real roots. There always exists a factorization of $g(x)$ to linear factors with eventual complex roots though. Repeating the previous notion for complex polynomials gives us the same result. If we know in advance the coefficients of the polynomials are real though, the complex roots in our expressions will come up simultaneously with their complex conjugate roots. Therefore we work with quadratic factors of the form of sum of squares $(x - a)^2 + b^2$ and their powers straight out. Our previous notion work very well again and guarantees that it will be possible to see the respective summands in the form of

$$\frac{Bx + C}{((x - a)^2 + b^2)^n}.$$

Similarly to the real roots case, we can always find the 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 (irreducble) factor as well. Specific results can also be tried out in Maple by calling the procedure "convert(h, parfrac, x)" that decomposes the expression h that is rationally dependant on the variable x into partial fractions.

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

$$\int f(\varphi(x))\,\varphi'(x)\,dx \tag{6.8}$$

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

$$\int \cos^5 x\cdot\sin x\,dx=-\int\cos^5 x\,(-\sin x)\;dx=-\int y^5\,dy=\\ -\tfrac{y^6}{6}+C=-\tfrac{\cos^6 x}{6}+C.$$

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

$$\int\cos^5 x\cdot\sin^2 x\,dx=\left|\begin{array}{c} t=\sin x\\ dt=\cos x\,dx\end{array}\right|=\int\left(1-t^2\right)^2 t^2\,dt=\\ \int t^6-2t^4+t^2\,dt=\tfrac{t^7}{7}-2\tfrac{t^5}{5}+\tfrac{t^3}{3}+C=\tfrac{\sin^7 x}{7}-\tfrac{2\sin^5 x}{5}+\tfrac{\sin^3 x}{3}+C.$$

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 (‖6.8‖) 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 such a form that we may just be able to compute it. We want to determine the primitive function of function $f(x)=\mathrm{tg}^4\,x$. Thus it's sensible to consider the substitution $u=\mathrm{tg}\,x$. We obtain

$$\int\tfrac{\sin^4 x}{\cos^4 x}\,dx=\left|\begin{array}{c} x=\mathrm{arctg}\,u\\ dx=\tfrac{du}{1+u^2}\end{array}\right|=\int\tfrac{u^4}{1+u^2}\,du=\int u^2-1+\tfrac{1}{u^2+1}\,du=\\ \tfrac{u^3}{3}-u+\mathrm{arctg}\,u+C=\tfrac{\mathrm{tg}^3\,x}{3}-\mathrm{tg}\,x+\mathrm{arctg}\,(\mathrm{tg}\,x)+C=\\ \tfrac{\mathrm{tg}^3\,x}{3}-\mathrm{tg}\,x+x+C.$$

Case (d). We have

$$\int\tfrac{1-\sqrt[3]{x}+\sqrt{x}}{\sqrt[6]{x^5}+x}\,dx=\left|\begin{array}{c} z^6=x\\ 6z^5\,dz=dx\end{array}\right|=\int\tfrac{1-z^2+z^3}{z^5+z^6}\,6z^5\,dz=\\ 6\int\tfrac{1-z^2+z^3}{1+z}\,dz=6\int z^2-2z+2-\tfrac{1}{z+1}\,dz=\\ 6\left(\tfrac{z^3}{3}-z^2+2z-\ln|\,z+1\,|\right)+C=\\ 2\sqrt{x}-6\sqrt[3]{x}+12\sqrt[6]{x}-6\ln\left(\sqrt[6]{x}+1\right)+C,$$

where we again easily determined by substitution (for $z\neq-1$)

$$\int\tfrac{dz}{z+1}=\left|\begin{array}{c} v=z+1\\ dv=dz\end{array}\right|=\int\tfrac{dv}{v}=\ln|\,v\,|+C=\ln|\,z+1\,|+C.$$

□

6.50. By combining integration by parts and the substitution method, determine

(a) $\int x^3\,\mathrm{e}^{-x^2}\,dx,\ x\in\mathbb{R};$

(b) $\int x\arcsin x^2\;dx,\ x\in(-1,1).$

Solution. Case (a). The substitution method leads to the integral

We can already integrate all of the above shown partial fractions. Recall that the last mentioned ones lead to integrals discussed in example 6.22 among other things.

To sum up, the rational function $f(x)/g(x)$ can be integrated fairly easily, if we can find the corresponding decomposition of the polynomial in the denominator $g(x)$. While computing Newton integrals though, the problematic points are the discontinuities of rational functions, in whose neighbourhood these functions are unbounded. We will tend to this problem later (see paragraph 6.30 lower).

6.24. Riemann integral. The notion of computing integral as a representation of the area bounded by the graph of a function and the x axis has to be put more precisely. We'll do that now and then prove that for all continuous function this definition yields the same results as Newton integral.

Consider a real function f defined on the interval $[a,b]$ and choose a partition of this interval along with the choice of representants ξ_i of the respective parts, i.e. $a=x_0<x_1<\cdots<x_n=b$ and $\xi_i\in[x_{i-1},x_i]$, $i=1,\dots,n$. The number $\delta=\min_i\{x_i-x_{i-1}\}$ is called the *norm of the partition*. We define *Riemann sum* corresponding to the chosen partition along with the chosen representants

$$\Xi=(x_0,\dots,x_n;\xi_1,\dots,\xi_n)$$

as

$$S_\Xi=\sum_{i=1}^{n}f(\xi_i)\cdot(x_i-x_{i-1}).$$

We say the *Riemann integral* of function f on the interval $[a,b]$ exists, if for every sequence of partitions with representants $(\Xi_k)_{k=0}^{\infty}$ with norms of the partitions δ_k approaching zero, the limit

$$\lim_{k\to\infty}S_{\Xi_k}=S$$

exists and its value doesn't depend on the choice of the sequence of partitions and their representants. In that case we write

$$S=\int_a^b f(x)dx.$$

This definition doesn't look too practically, but nonetheless it will allow us to easily formulate and prove several simple properties of the Riemann integral:

Theorem. *(1) If 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*

$$\int_a^b f(x)dx=\int_a^c f(x)dx+\int_c^b f(x)dx$$

holds.

(2) If f and g are two real functions defined on the interval $[a,b]$ and 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

$$\int_a^b(f(x)+g(x))dx=\int_a^b f(x)dx+\int_a^b g(x)dx$$

holds.

$$\int x^3\,\mathrm{e}^{-x^2}\,dx = \left|\begin{array}{c} t=-x^2 \\ dt=-2x\,dx \end{array}\right| = \tfrac12\int t\,\mathrm{e}^t\,dt,$$

which can be easily computed by integrating by parts, yielding

$$\tfrac12\int t\,\mathrm{e}^t\,dt = \left|\begin{array}{c|c} F(t)=t & F'(t)=1 \\ G'(t)=\mathrm{e}^t & G(t)=\mathrm{e}^t \end{array}\right| = \tfrac12 t\,\mathrm{e}^t - \tfrac12\int \mathrm{e}^t\,dt = \tfrac12 t\,\mathrm{e}^t - \tfrac12\,\mathrm{e}^t + C = -\tfrac12\,\mathrm{e}^{-x^2}\left(x^2+1\right)+C.$$

Case (b). Similarly, we obtain

$$\int x\arcsin x^2\,dx = \left|\begin{array}{c} t=x^2 \\ dt=2x\,dx \end{array}\right| = \tfrac12\int \arcsin t\,dt = \left|\begin{array}{c|c} F(t)=\arcsin t & F'(t)=\frac{1}{\sqrt{1-t^2}} \\ G'(t)=1 & G(t)=t \end{array}\right| = \tfrac12 t\arcsin t - \tfrac12\int\tfrac{t}{\sqrt{1-t^2}}\,dt =$$

$$\left|\begin{array}{c} u=1-t^2 \\ du=-2t\,dt \end{array}\right| = \tfrac12 t\arcsin t + \tfrac14\int\tfrac{du}{\sqrt u} = \tfrac12 t\arcsin t + \tfrac12\sqrt{u}+C = \tfrac12 t\arcsin t + \tfrac12\sqrt{1-t^2}+C = \tfrac12 x^2\arcsin x^2 + \tfrac12\sqrt{1-x^4}+C.$$

□

6.51. Compute the integral

$$\int\sqrt{1-x^2}\,dx, \qquad x\in(-1,1)$$

in two different ways.

Solution. Integration by parts yields

$$\int\sqrt{1-x^2}\,dx = \left|\begin{array}{c|c} F(x)=\sqrt{1-x^2} & F'(x)=\frac{-x}{\sqrt{1-x^2}} \\ G'(x)=1 & G(x)=x \end{array}\right| = x\sqrt{1-x^2}+\int\tfrac{x^2}{\sqrt{1-x^2}}\,dx = x\sqrt{1-x^2}-\int\tfrac{1-x^2-1}{\sqrt{1-x^2}}\,dx =$$

$$x\sqrt{1-x^2}-\int\sqrt{1-x^2}\,dx+\int\tfrac{1}{\sqrt{1-x^2}}\,dx = x\sqrt{1-x^2}-\int\sqrt{1-x^2}\,dx+\arcsin x,$$

which implies

$$2\int\sqrt{1-x^2}\,dx = x\sqrt{1-x^2}+\arcsin x + C,$$

i.e.

$$\int\sqrt{1-x^2}\,dx = \tfrac12\left(x\sqrt{1-x^2}+\arcsin x\right)+C.$$

The substitution method along with (‖6.7‖) then yields

$$\int\sqrt{1-x^2}\,dx = \left|\begin{array}{c} x=\sin y \\ dx=\cos y\,dy \end{array}\right| = \int\sqrt{1-\sin^2 y}\cdot\cos y\,dy = \int\cos^2 y\,dy = \tfrac12\,(y+\sin y\cdot\cos y)+C =$$

$$\tfrac12\left(\sin y\cdot\sqrt{1-\sin^2 y}+y\right)+C = \tfrac12\left(x\sqrt{1-x^2}+\arcsin x\right)+C,$$

where $y\in(-\pi/2,\pi/2)$ for $x\in(-1,1)$, thus among other things, we have

$$0<\cos y = |\cos y| = \sqrt{\cos^2 y} = \sqrt{1-\sin^2 y}.$$

□

6.52. Determine

$$\int \mathrm{e}^{\sqrt{x}}\,dx, \qquad x>0.$$

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

(3) If f is 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

$$\int_a^b C\cdot f(x)dx = C\cdot\int_a^b f(x)dx$$

holds.

Proof. (1) First suppose that the integral over the whole interval exists. When computing it, surely 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 representants in limit, then the total sums couldn't be independant on the choices in limit (it suffices to keep of sequence of partition of the subinerval the same and change the other so 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 independantly on their choice. If we add a partitioning point c to any sequence of Riemann sums over the whole interval $[a,b]$, we'll change the value of the whole sum and also the values of partial sums over the intervals belonging to $[a,c]$ and $[c,b]$ at most by a multiple of the norm of the partition and possible differences of the bounded function f on whole $[a,b]$. That's a number tending arbitrarily close to zero for a decreasing norm of the partition. Then necessarily the partial Riemann sums of our function 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 representants. Because multiplication of real numbers is distributive, the statement follows.

(3) The same thought as in the previous case. □

The following result is crucial for understanding the relation between an integral and a derivative:

6.25. Theorem (The fundamental theorem of integral calculus)**.** *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 $F(t)$ 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.

The whole proof of this important statement will be somewhat longer. In the first step for proving the existence of the integral, we'll use an alternative definition, in which we replace the choice of representants and the corresponding value $f(\xi_i)$ by the suprema M_i of the values $f(x)$ in the corresponding subinterval $[x_{i-1},x_i]$, or by theinfima m_i of the function $f(x)$ in the same subinterval, respectively. We speak of upper and lower Riemann sums, respectively (in literature, this process is sometimes called the *Darboux integral*).

$$\int \mathrm{e}^{\sqrt{x}}\,dx = \left|\begin{matrix} y^2 = x \\ 2y\,dy = dx \end{matrix}\right| = 2\int y\,\mathrm{e}^y\,dy.$$

Now by using integration by parts, we'll compute

$$\int y\,\mathrm{e}^y\,dy = \left|\begin{matrix} F(y) = y & F'(y) = 1 \\ G'(y) = \mathrm{e}^y & G(y) = \mathrm{e}^y \end{matrix}\right| = y\,\mathrm{e}^y - \int \mathrm{e}^y\,dy = y\,\mathrm{e}^y - \mathrm{e}^y + C.$$

Thus in total, we have

$$\int \mathrm{e}^{\sqrt{x}}\,dx = 2y\,\mathrm{e}^y - 2\,\mathrm{e}^y + C = 2\,\mathrm{e}^{\sqrt{x}}\left(\sqrt{x} - 1\right) + C.$$

□

6.53. 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. □

C. Integration of rational functions

6.54. 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

$$\int \tfrac{6}{x-2}\,dx = \left|\begin{matrix} y = x-2 \\ dy = dx \end{matrix}\right| = \int \tfrac{6}{y}\,dy = 6\ln|\,y\,| + C = 6\ln|\,x-2\,| + C$$

and similarly

$$\int \tfrac{6}{(x+4)^3}\,dx = \left|\begin{matrix} y = x+4 \\ dy = dx \end{matrix}\right| = \int \tfrac{6}{y^3}\,dy = \tfrac{6}{-2y^2} + C = -\tfrac{3}{(x+4)^2} + C.$$

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

$$\int \tfrac{A}{x-x_0}\,dx = \left|\begin{matrix} y = x-x_0 \\ dy = dx \end{matrix}\right| = \int \tfrac{A}{y}\,dy = A\ln|\,y\,| + C = A\ln|\,x-x_0\,| + C$$

and

$$\int \tfrac{A}{(x-x_0)^n}\,dx = \left|\begin{matrix} y = x-x_0 \\ dy = dx \end{matrix}\right| = \int \tfrac{A}{y^n}\,dy = \tfrac{Ay^{-n+1}}{-n+1} + C = \tfrac{A}{(1-n)(x-x_0)^{n-1}} + C$$

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 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

6.26. Upper and lower Riemann integral. Because our function is continuous, it's surely bounded on a closed interval, hence all the above considered suprema and infima and finite. Then the *upper Riemann sum* corresponging to the partition $\Xi = (x_0, \dots, x_n)$ is given by the expression

$$S_{\Xi,\sup} = \sum_{i=1}^{n}\Big(\sup_{x_{i-1}\leq\xi\leq x_i} f(\xi)\Big)\cdot(x_i - x_{i-1}) = \sum_{i=1}^{n} M_i(x_i - x_{i-1}),$$

while the *lower Riemann sum* is

$$S_{\Xi,\inf} = \sum_{i=1}^{n}\Big(\inf_{x_{i-1}\leq\xi\leq x_i} f(\xi)\Big)\cdot(x_i - x_{i-1}) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}).$$

Because for every partition $\Xi = (x_0, \dots, x_n; \xi_1, \dots, \xi_n)$ with representants, the inequalities

$$S_{\Xi,\inf} \leq S_{\Xi,\xi} \leq S_{\Xi,\sup} \tag{6.3}$$

hold, and the infima and suprema can be approximated with arbitrary precision by the real values, we can suspect that the Riemann integral will exists if and only if for every sequences of partitions with norm approaching zero, the limits of both the upper and lower sums will exists and they will be equal. We'll prove that 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 a bounded function f over the interval $[a, b]$ exists iff $S_{\sup} = S_{\inf}$.

Proof. If we refine a partition Ξ_1 to Ξ_2 by adding another points, clearly

$$S_{\Xi_1,\sup} \geq S_{\Xi_2,\sup}, \qquad S_{\Xi_1,\inf} \leq S_{\Xi_2,\inf}.$$

Every two partitions have common refinement, 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. Indeed, if the common limit of the upper sums S exists and is independant on the chosen sequence of partitions, then it must be $S_{\sup}$, and similarly for lower sums.

Conversely, consider a fixed partition Ξ with n inner partitioning points of the intervala $[a, b]$, and another partition Ξ_1, whose norm is a very small number δ. In their common refinement Ξ_2, there will be only n intervals, that will contribute to the sum $S_{\Xi_2,\sup}$ by eventually lesser contribution than in the case of Ξ_1. Because f is a bounded function on$[a, b]$, each of these contributions will be bounded by a universal constant multiplied by the norm of the partition (i.e. the maximal length of the corresponding interval in the partition). Hence when choosing sufficiently small δ, the distance of $S_{\Xi_1,\sup}$ from $S_{\sup}$ won't be bigger than twice the distance of $S_{\Xi,\sup}$ from $S_{\sup}$.

If we now choose an arbitrary sequence Ξ_k with upper sums, whose limit is $S_{\sup}$, then for fixed $\varepsilon > 0$ we can find k such that $S_{\Xi_k,\sup}$, $k \geq N$ will be distant from $S_{\sup}$ by less than ε.

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

$$\int \frac{3x+7}{x^2-4x+15}\,dx = \frac{3}{2}\int \frac{2x-4}{x^2-4x+15}\,dx + 13\int \frac{dx}{x^2-4x+15} =$$
$$\frac{3}{2}\ln\left(x^2-4x+15\right) + 13\int \frac{dx}{(x-2)^2+11} = \frac{3}{2}\ln\left(x^2-4x+15\right) +$$
$$\frac{13}{11}\int \frac{dx}{\left(\frac{x-2}{\sqrt{11}}\right)^2+1} = \left|\begin{array}{l} y = \frac{x-2}{\sqrt{11}} \\ dy = \frac{dx}{\sqrt{11}} \end{array}\right| = \frac{3}{2}\ln\left(x^2-4x+15\right) + \frac{13}{\sqrt{11}}\int \frac{dy}{y^2+1} =$$
$$\frac{3}{2}\ln\left(x^2-4x+15\right) + \frac{13}{\sqrt{11}}\operatorname{arctg} y + C =$$
$$\frac{3}{2}\ln\left(x^2-4x+15\right) + \frac{13}{\sqrt{11}}\operatorname{arctg}\frac{x-2}{\sqrt{11}} + C.$$

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

$$\int \frac{2(x-x_0)}{(x-x_0)^2+a^2}\,dx = \left|\begin{array}{l} y = (x-x_0)^2+a^2 \\ dy = 2\,(x-x_0)\,dx \end{array}\right| = \int \frac{dy}{y} =$$
$$\ln|\,y\,| + C = \ln[(x-x_0)^2+a^2] + C,$$
$$\int \frac{1}{(x-x_0)^2+a^2}\,dx = \frac{1}{a^2}\int \frac{dx}{\left(\frac{x-x_0}{a}\right)^2+1} = \left|\begin{array}{l} z = \frac{x-x_0}{a} \\ dz = \frac{dx}{a} \end{array}\right| = \frac{1}{a}\int \frac{dz}{z^2+1} =$$
$$\frac{1}{a}\operatorname{arctg} z + C = \frac{1}{a}\operatorname{arctg}\frac{x-x_0}{a} + C,$$

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

$$\frac{Ax+B}{\left[(x-x_0)^2+a^2\right]^n}, \qquad A, B, x_0 \in \mathbb{R},\ a > 0, n \in \mathbb{N} \setminus \{1\},$$

which can be analogically simplified to

$$\frac{A}{2}\cdot\frac{2(x-x_0)}{\left[(x-x_0)^2+a^2\right]^n} + (B+Ax_0)\cdot\frac{1}{\left[(x-x_0)^2+a^2\right]^n}.$$

Then we'll determine

$$\int \frac{2(x-x_0)}{\left[(x-x_0)^2+a^2\right]^n}\,dx = \left|\begin{array}{l} y = (x-x_0)^2+a^2 \\ dy = 2\,(x-x_0)\,dx \end{array}\right| = \int \frac{dy}{y^n} =$$
$$\frac{1}{(1-n)y^{n-1}} + C = \frac{1}{(1-n)\left[(x-x_0)^2+a^2\right]^{n-1}} + C$$

and

$$K_n\,(x_0,a) := \int \frac{1}{\left[(x-x_0)^2+a^2\right]^n}\,dx =$$
$$\left|\begin{array}{l|l} F(x) = \frac{1}{\left[(x-x_0)^2+a^2\right]^n} & F'(x) = \frac{-2n(x-x_0)}{\left[(x-x_0)^2+a^2\right]^{n+1}} \\ G'(x) = 1 & G(x) = x - x_0 \end{array}\right| =$$
$$\frac{x-x_0}{\left[(x-x_0)^2+a^2\right]^n} + 2n\int \frac{(x-x_0)^2+a^2}{\left[(x-x_0)^2+a^2\right]^{n+1}} - \frac{a^2}{\left[(x-x_0)^2+a^2\right]^{n+1}}\,dx =$$
$$\frac{x-x_0}{\left[(x-x_0)^2+a^2\right]^n} + 2n\left(K_n\,(x_0,a) - a^2\,K_{n+1}\,(x_0,a)\right),$$

which implies

$$K_{n+1}\,(x_0,a) = \frac{1}{a^2}\left(\frac{2n-1}{2n}\,K_n\,(x_0,a) + \frac{1}{2n}\frac{x-x_0}{\left[(x-x_0)^2+a^2\right]^n}\right),$$

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

$$\int \frac{30x-77}{\left(x^2-6x+13\right)^2}\,dx = 15\int \frac{2x-6}{\left(x^2-6x+13\right)^2}\,dx + 13\int \frac{1}{\left(x^2-6x+13\right)^2}\,dx$$

and furthermore

According to the last notion, we can find δ such that for all partitions with norm less than δ the sum will be closer than 2ε. Hence we have just shown that for arbitrary $\varepsilon > 0$, we can find $\delta > 0$ such that for all partitions with norm at most δ the inequality $|S_{\Xi,\sup} - S_\Xi| < \varepsilon$ will hold. That's exactly the statement that the number $S_{\sup}$ is the limit of all sequences of upper sums with norms of the partition approaching zero. We cam prove the statement for lower sums in exactly the same way.

If the Riemann integral doesn't exist, there exist sequences of partitions and representants with different limits of Riemann sums. The the proven statement implies, that the limits of upper and lower sums will be different as well. Conversely, suppose $S_{\sup} = S_{\inf}$, but then all Riemann sums of sequences of the partitions must have the same limit because of the inequalities (6.3). □

6.27. Uniform continuity. Until now, we have only used the continuity of our function f to show that all such functions are bounded on a closed finite interval. We still have to show though, that for continuous functions, $S_{\sup} = S_{\inf}$.

>From the definition of continuity we know that 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 also be rewritten in this way: for $y, z \in \mathcal{O}_\delta(x)$, i.e.

$$|y - z| < 2\delta,$$

it's true that $f(y), f(z) \in \mathcal{O}_\varepsilon(f(x))$, i.e.

$$|f(y) - f(z)| < 2\varepsilon.$$

We're going to need a global variant of such property, we call it*uniform continuity* of function f:

Theorem. *Let f be a continuous function on a closed finite interval $[a,b]$. Then 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.*

Proof. Because every finite closed interval is compact, we can cover it by finitely many neighbourhoods $\mathcal{O}_{\delta(x)}(x)$ mentioned above in connection with continuity. Their radius $\delta(x)$ depends on the center x, while we'll consider the number ε fixed. Finally, we'll choose δ as the minimum of all (finitely many) $\delta(x)$. Our continuous function f then has the desired property (we only interchange the numbers ε and δ for their doubles). □

6.28. Finishing the proof of Theorem 6.25. Now we can easily finish the whole proof of existence of the Riemann integral. Choose ε and δ the same way as in the previous theorem about uniform continuity and consider any partition Ξ with n intervals and norm at most δ. Then

$$\left|\sum_{i=1}^{n} \sup_{x_{i-1}\le\xi\le x_i} f(\xi)\cdot(x_i - x_{i-1}) - \sum_{i=1}^{n} \inf_{x_{i-1}\le\xi\le x_i} f(\xi)\cdot(x_i - x_{i-1})\right|$$
$$\le \sum_{i=1}^{n}\left|\sup_{x_{i-1}\le\xi\le x_i} f(\xi) - \inf_{x_{i-1}\le\xi\le x_i} f(\xi)\right|\cdot(x_i - x_{i-1})$$
$$\le \varepsilon\cdot(b-a).$$

$$\int \frac{2x-6}{(x^2-6x+13)^2}\,dx = \left| \begin{array}{c} y = x^2-6x+13 \\ dy = (2x-6)\,dx \end{array} \right| = \int \frac{dy}{y^2} = -\frac{1}{y} + C = -\frac{1}{x^2-6x+13} + C,$$

$$\int \frac{1}{(x^2-6x+13)^2}\,dx = \int \frac{dx}{\left[(x-3)^2+2^2\right]^2} = \frac{1}{2^2}\left(\frac{2-1}{2}\,K_1(3,2) + \frac{1}{2}\,\frac{x-3}{(x-3)^2+2^2}\right) =$$
$$\frac{1}{4}\left(\frac{1}{4}\operatorname{arctg}\frac{x-3}{2} + C + \frac{1}{2}\,\frac{x-3}{x^2-6x+13}\right) = \frac{1}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{1}{8}\,\frac{x-3}{x^2-6x+13} + C.$$

In total, we have

$$\int \frac{30x-77}{(x^2-6x+13)^2}\,dx = -\frac{15}{x^2-6x+13} + \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{13}{8}\,\frac{x-3}{x^2-6x+13} + C = \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{13x-159}{8(x^2-6x+13)} + C.$$

□

6.55. 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)

$$\int \frac{x-4}{5x^2+6x+3}\,dx = \frac{1}{10}\int \frac{10x+6}{5x^2+6x+3}\,dx - \frac{23}{5}\int \frac{dx}{5x^2+6x+3} = \frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23}{25}\int \frac{dx}{\left(x+\frac{3}{5}\right)^2+\frac{6}{25}} =$$
$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23}{6}\int \frac{dx}{\left(\frac{5x+3}{\sqrt{6}}\right)^2+1} = \left| \begin{array}{c} t = \frac{5x+3}{\sqrt{6}} \\ dt = \frac{5}{\sqrt{6}}\,dx \end{array} \right| =$$
$$\frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt{6}}{30}\int \frac{dt}{t^2+1} = \frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt{6}}{30}\operatorname{arctg} t + C = \frac{1}{10}\ln\left(5x^2+6x+3\right) - \frac{23\sqrt{6}}{30}\operatorname{arctg}\frac{5x+3}{\sqrt{6}} + C;$$

(c)

$$\int \frac{dx}{(x-4)(x-2)(x^2+2x+2)} = \frac{1}{52}\int \frac{dx}{x-4} - \frac{1}{20}\int \frac{dx}{x-2} + \frac{1}{130}\int \frac{4x+11}{x^2+2x+2}\,dx = \frac{1}{52}\ln|\,x-4\,| - \frac{1}{20}\ln|\,x-2\,| + \frac{1}{130}\left(2\int \frac{2x+2}{x^2+2x+2}\,dx + 7\int \frac{dx}{x^2+2x+2}\right) =$$
$$\frac{1}{260}\ln\left|\frac{(x-4)^5}{(x-2)^{13}}\right| + \frac{2}{130}\ln\left(x^2+2x+2\right) + \frac{7}{130}\int \frac{dx}{(x+1)^2+1} = \left| \begin{array}{c} t = x+1 \\ dt = dx \end{array} \right| = \frac{1}{260}\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + \frac{7}{130}\int \frac{dt}{t^2+1} =$$

We can see that for decreasing norm of the partition the upper and lower sums are arbitrarily close to each other. That's why the suprema and infima coincide, which is what we needed to show.

For a complete proof of the fundamental theorem of integral calculus, we still need to verify the statement about the existence of the primitive function. We already know that for continuous function f on interval $[a,b]$ there exists the integral $\int_a^t f(x)dx$ for every $t \in [a,b]$. As in the statement about uniform continuity, we can choose $\delta > 0$, dependant on dixed small $\varepsilon > 0$, such that

$$|f(x+\Delta x) - f(x)| < \varepsilon$$

for all $0 \leq \Delta x < \delta$ on the whole intervalu$[a,b]$. The difference of the derivative of our function $F(x)$ and the integrated function $f(x)$ is expressed by the limit of the expressions

$$\frac{1}{\Delta x}\left(\int_a^{x+\Delta x} f(t)dt - \int_a^t f(t)dt\right) - f(x)$$

for Δx approaching zero. If we choose $0 < \Delta x < \delta$ though, then in absolute value this expression can be estimated by

$$\left|\frac{1}{\Delta x}\left(\int_t^{x+\Delta x} f(t)dt\right) - f(x)\right| < \varepsilon,$$

because in the expression on the left hand side we can replace the integral by its Riemann sum with arbitrary precision and in the summands $f(\xi_i)(x_i - x_{i-1})$ with $\xi_i \in [x, x+\Delta x]$ in any Riemann sum, the values $f(\xi)$ are distant from $f(x)$ by at most ε. Hence by replacing all $f(\xi_i)$ by $f(x)$, we obtain a zero expression on the left hand side with error of at most ε.

But that means that at the point x, the right derivative of function $F(x)$ exists and equals $f(x)$. We prove the result for the left derivative in the same way, and the whole theorem 6.25 is therefore proven.

6.29. Important notes. (1) Theorems 6.25 and 6.24 claim that integral is a linear map

$$\int : C[a,b] \to \mathbb{R}$$

from a vector space of continuous functions on interval $[a,b]$ to real numbers. Hence it's 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 Newton and Riemann integral 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 6.25 we also proved an important statement, that for bounded function f on interval $[a,b]$, the limits of the upper and lower sums always exist. They are also called *the upper Riemann integral* and *the lower Riemann integral* and they are often denoted by $\overline{\int}_a^b f(x)\,dx$ and $\underline{\int}_a^b f(x)\,dx$.

In this way we can equivalently define the Riemann integral for continuous functions (as we did in the proof).

(4) In the next step of the proof we derived an important property of continuous functions that is called *the uniform continuity* on a closed interval $[a,b]$. Clearly every uniform continuous function is continuous as well, but the converse need not be true on open intervals. As an example, consider the function $f(x) = \sin(1/x)$ on the interval $(0,1)$.

$$\frac{1}{260}\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + \frac{7}{130}\operatorname{arctg} t + C =$$
$$\frac{1}{260}\left[\ln\left|\frac{(x-4)^5\left(x^2+2x+2\right)^4}{(x-2)^{13}}\right| + 14\operatorname{arctg}(x+1)\right] + C;$$

(d)

$$\int\frac{x}{x^4-x^3-x+1}\,dx = \frac{1}{3}\int\frac{dx}{(x-1)^2} - \frac{1}{3}\int\frac{dx}{x^2+x+1} =$$
$$-\frac{1}{3(x-1)} - \frac{1}{3}\int\frac{dx}{\left(x+\frac{1}{2}\right)^2+\frac{3}{4}} = -\frac{1}{3(x-1)} - \frac{4}{9}\int\frac{dx}{\left(\frac{2x+1}{\sqrt{3}}\right)^2+1} =$$
$$\left|\begin{array}{c} t = \frac{2x+1}{\sqrt{3}} \\ dt = \frac{2}{\sqrt{3}}\,dx \end{array}\right| = -\frac{1}{3(x-1)} - \frac{2}{3\sqrt{3}}\int\frac{dt}{t^2+1} =$$
$$-\frac{1}{3(x-1)} - \frac{2}{3\sqrt{3}}\operatorname{arctg} t + C = -\frac{1}{3(x-1)} - \frac{2}{3\sqrt{3}}\operatorname{arctg}\frac{2x+1}{\sqrt{3}} + C;$$

(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)

$$\int\frac{5x^2-12}{x^4-12x^3+62x^2-156x+169}\,dx = \int\frac{5x^2-12}{(x^2-6x+13)^2}\,dx =$$
$$5\int\frac{dx}{x^2-6x+13} + \int\frac{30x-77}{(x^2-6x+13)^2}\,dx =$$
$$5\int\frac{dx}{(x-3)^2+4} + 15\int\frac{2x-6}{(x^2-6x+13)^2}\,dx + 13\int\frac{dx}{(x^2-6x+13)^2} =$$
$$\frac{5}{4}\int\frac{dx}{\left(\frac{x-3}{2}\right)^2+1} + 15\int\frac{2x-6}{(x^2-6x+13)^2}\,dx + 13\int\frac{dx}{\left[(x-3)^2+4\right]^2} =$$
$$\left|\begin{array}{c} t = \frac{x-3}{2} \\ dt = \frac{1}{2}\,dx \end{array}\right| \left.\begin{array}{c} u = x^2-6x+13 \\ du = (2x-6)\,dx \end{array}\right| = \frac{5}{2}\int\frac{dt}{t^2+1} + 15\int\frac{du}{u^2} +$$
$$\frac{13}{16}\int\frac{dx}{\left[\left(\frac{x-3}{2}\right)^2+1\right]^2} = \frac{5}{2}\operatorname{arctg} t - \frac{15}{u} + \frac{13}{8}\int\frac{dt}{\left[t^2+1\right]^2} =$$
$$\frac{5}{2}\operatorname{arctg}\frac{x-3}{2} - \frac{15}{x^2-6x+13} + \frac{13}{8}\left(\frac{1}{2}\operatorname{arctg} t + \frac{1}{2}\,\frac{t}{t^2+1}\right) + C =$$
$$\frac{5}{2}\operatorname{arctg}\frac{x-3}{2} - \frac{15}{x^2-6x+13} + \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{13}{16}\,\frac{\frac{x-3}{2}}{\left(\frac{x-3}{2}\right)^2+1} + C =$$
$$\frac{5}{2}\operatorname{arctg}\frac{x-3}{2} + \frac{13}{16}\operatorname{arctg}\frac{x-3}{2} - \frac{15}{x^2-6x+13} + \frac{13}{8}\,\frac{x-3}{(x-3)^2+4} + C =$$
$$\frac{53}{16}\operatorname{arctg}\frac{x-3}{2} + \frac{13x-159}{8(x^2-6x+13)} + C.$$

□

6.56. Compute

$$\int\frac{x}{(x-1)^2\left(x^2+2x+2\right)}\,dx, \qquad x \neq 1.$$

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

$$\frac{x}{(x-1)^2\left(x^2+2x+2\right)} = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{Cx+D}{x^2+2x+2}$$

(5) Consider a function f on an interval $[a, b]$, which is only *sequentially continuous*. That means it's 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 integral with respect to the interval of integration, see 6.24(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 function $F(x)$ exists in all points x, in which f is continuous. Moreover it can easily be verified that $F(x)$ is continuous at the remaining points, so it's a continuous function on the whole interval $[a, b]$. When evaluating the integral by primitive functions, it's necessary to choose its individual parts so that they are connected. Then the whole integral can be computed as a difference of the function $F(x)$ in its boundary values.

(6) The Lagrange theorem of mean value of a differentiable function has an analogy that is called *the integral mean value theorem*. Consider a function$f(x)$ that is continuous on an interval $[a, b]$ and its primitive function $F(x)$. The mean value theorem claims that there exists an inner point $a < c < b$ such that

$$\int_a^b f(x)\,dx = F(b) - F(a) = F'(c)(b-a) = f(c)(b-a).$$

This statement can be fairly easily derived directly from the definition of the Riemann integral and then it can be used in a straightforward way in the final step of the proof of the fundamental theorem of integral calculus.

6.30. Improper integrals. When discussing integration of rational functions, we saw, that we would also like to work with definite integrals over intervals, where the integrated function $f(x)$ has improper (one-sided) limits. In that case, the integrated function is neither continuous nor bounded and thus may not satisfy the earlier derived results. We speak of the "improper integral".

A simple help in this case is discussing the definite integrals on smaller intervals with the boundary approaching the problematic point and study, whether the limit value of such definite integrals exists. If it does, we say the corresponding improper integral exists and equals this limit. We'll show this procedure on a simple example:

$$I = \int_0^2 \frac{dx}{\sqrt[4]{2-x}}$$

is an improper integral, because the integrated function $f(x) = (2-x)^{-1/4}$ has its left-sided limit at the point $b = 2$ which equals ∞. The integrated function is continuous at all other points. Therefore we study the integrals

$$I_\delta = \int_0^{2-\delta}\frac{dx}{\sqrt[4]{2-x}} = \int_\delta^2 y^{-1/4}\,dy$$
$$= \left[-\frac{4}{3}y^{3/4}\right]_\delta^2 = \frac{4}{3}2^{3/4} - \frac{4}{3}\delta^{3/4}.$$

Notice that when we used substitution, we obtained an integral with recalculated upper bound δ and lower bound 2. By transforming the bounds to a usual position, we add one minus sign.

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 $x \in \mathbb{R} \setminus \{1\}$. 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

$$x - \tfrac{1}{5}\left(x^2+2x+2\right) = -\tfrac{1}{5}x^2 + \tfrac{3}{5}x - \tfrac{2}{5},$$
$$A\,(x-1)\left(x^2+2x+2\right) + (Cx+D)\,(x-1)^2 =$$
$$(A+C)\,x^3 + (A-2C+D)\,x^2 + (C-2D)\,x - 2A + D,$$

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

$$\int \tfrac{x}{(x-1)^2\left(x^2+2x+2\right)}\,dx = \int \tfrac{dx}{25(x-1)} + \int \tfrac{dx}{5(x-1)^2} - \int \tfrac{x+8}{25\left(x^2+2x+2\right)}\,dx =$$
$$\tfrac{1}{25}\ln|\,x-1\,| - \tfrac{1}{5(x-1)} - \tfrac{1}{50}\ln\left(x^2+2x+2\right) - \tfrac{7}{25}\operatorname{arctg}(x+1) + C,$$

where we used

$$\int \tfrac{x+8}{x^2+2x+2}\,dx = \int \tfrac{\frac{1}{2}(2x+2)}{x^2+2x+2} + \tfrac{7}{x^2+2x+2}\,dx = \tfrac{1}{2}\int \tfrac{2x+2}{x^2+2x+2}\,dx +$$
$$7\int \tfrac{1}{(x+1)^2+1}\,dx = \tfrac{1}{2}\ln\left(x^2+2x+2\right) + 7\operatorname{arctg}(x+1) + C.$$

□

6.57. 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

$$\left(x^3+2x^2+x-1\right) : \left(x^2-x+1\right) = x + 3 + \tfrac{3x-4}{x^2-x+1},$$

to consider a proper rational function (with the degree of the numerator lower then the degree of the denominator). Now we'll compute

$$\int \tfrac{x^3+2x^2+x-1}{x^2-x+1}\,dx = \int x+3\,dx + \int \tfrac{3x-4}{x^2-x+1}\,dx =$$
$$\tfrac{x^2}{2} + 3x + \tfrac{3}{2}\int \tfrac{2x-1}{x^2-x+1}\,dx - \tfrac{5}{2}\int \tfrac{dx}{\left(x-\frac{1}{2}\right)^2+\left(\frac{\sqrt{3}}{2}\right)^2} =$$
$$\tfrac{x^2}{2} + 3x + \tfrac{3}{2}\ln\left(x^2-x+1\right) - \tfrac{5}{\sqrt{3}}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} + C.$$

Case (b). We have

$$\int \tfrac{x^8}{x^8-1}\,dx = \int 1\,dx + \tfrac{1}{8}\int \tfrac{dx}{x-1} - \tfrac{1}{8}\int \tfrac{dx}{x+1} - \tfrac{1}{4}\int \tfrac{dx}{x^2+1} +$$
$$\tfrac{1}{8}\int \tfrac{\sqrt{2}x-2}{x^2-\sqrt{2}x+1}\,dx - \tfrac{1}{8}\int \tfrac{\sqrt{2}x+2}{x^2+\sqrt{2}x+1}\,dx =$$

The limit for $\delta \to 0$ from the right clearly exists, so we've evaluated the improper integral

$$I = \int_0^2 \frac{dx}{\sqrt[4]{2-x}} = \frac{4}{3}2^{3/4}.$$

We can proceed in the same way if we want to integrated over an unbounded interval. In this case, we often speak of *improper integrals of the first kind*, while the integrals of unbounded functions on finite intervals are *improper integrals of the second kind.*

More generally, for example for $a \in \mathbb{R}$

$$I = \int_a^\infty f(x)\,dx = \lim_{b\to\infty}\int_a^b f(x)\,dx,$$

if the limit on the right hand side exists. Similarly we can have a finite upper bound and the other one infinite. If both are infinite, we evaluate the integral as a sum of two integrals with a chosen fixed bound in the middle, i.e.

$$\int_{-\infty}^\infty f(x)\,dx = \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx.$$

The existence nor the value doesn't depend on the choice of such bound, because by changing it, we only change both summands by the same finite value, but with the opposite sign. Conversely a limit for which the upper and lower bound would approach ±∞ 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 $\int_a^\infty x\,dx$ with one fixed bound approach infinite values fast.

When evaluating the improper integral of a rational function we must carefully divide the given interval according to the discontinuities of the integrated function and compute all the improper integrals seperately. Moreover it's necesarry to divide the whole interval in a way that we always integrate a function unbounded only in a neighbourhood of one of the boundary points.

6.31. New acquisitions to the ZOO. From the solved problems it could seem it's usual to find an indefinite integral by expressions composed of known elementary functions. That's a completely false impression.

On the contrary, an overwhelming majority of continuous functions leads to integrals we cannot express in this way. Even if we integrate fairly simple functions. Because the functions obtained by integration often appear in applications, many of them have names and before the advent of computers, extensive tables were published for the needs of engineers. In further text, we'll come back to the methods of obtaining numeric approximations of such functions.

Let's see at least some examples. In methods of signal processing, the function

$$\operatorname{sinc}(x) = \frac{\sin(x)}{x}$$

is very important. It can be verified in a way fairly straightforward, although toilful way, that it's a smooth function with limit values

$$f(0) = 1,\ f'(0) = 0,\ f''(0) = -\frac{2}{3}.$$

$$x + \tfrac{1}{8}\ln|\,x-1\,| - \tfrac{1}{8}\ln|\,x+1\,| - \tfrac{1}{4}\operatorname{arctg} x + \tfrac{\sqrt{2}}{16}\int \tfrac{2x-\sqrt{2}}{x^2-\sqrt{2}x+1}\,dx -$$
$$\tfrac{1}{8}\int \tfrac{dx}{\left(x-\frac{\sqrt{2}}{2}\right)^2+\left(\frac{\sqrt{2}}{2}\right)^2} - \tfrac{\sqrt{2}}{16}\int \tfrac{2x+\sqrt{2}}{x^2+\sqrt{2}x+1}\,dx - \tfrac{1}{8}\int \tfrac{dx}{\left(x+\frac{\sqrt{2}}{2}\right)^2+\left(\frac{\sqrt{2}}{2}\right)^2} =$$
$$x + \tfrac{1}{8}\ln|\,x-1\,| - \tfrac{1}{8}\ln|\,x+1\,| - \tfrac{1}{4}\operatorname{arctg} x +$$
$$\tfrac{\sqrt{2}}{16}\ln\left(x^2-\sqrt{2}x+1\right) - \tfrac{\sqrt{2}}{8}\operatorname{arctg}\left(\sqrt{2}x-1\right) -$$
$$\tfrac{\sqrt{2}}{16}\ln\left(x^2+\sqrt{2}x+1\right) - \tfrac{\sqrt{2}}{8}\operatorname{arctg}\left(\sqrt{2}x+1\right) + C.$$

□

6.58. Compute

$$\int \tfrac{2x^4+2x^2-5x+1}{x\left(x^2-x+1\right)^2}\,dx, \qquad x \neq 0.$$

Solution. We have

$$\int \tfrac{2x^4+2x^2-5x+1}{x\left(x^2-x+1\right)^2}\,dx = \int \tfrac{dx}{x} + \int \tfrac{x+3}{x^2-x+1}\,dx + \int \tfrac{x-6}{\left(x^2-x+1\right)^2}\,dx =$$
$$\ln|\,x\,| + \tfrac{1}{2}\int \tfrac{2x-1}{x^2-x+1}\,dx + \tfrac{7}{2}\int \tfrac{dx}{x^2-x+1} + \tfrac{1}{2}\int \tfrac{2x-1}{(x^2-x+1)^2}\,dx -$$
$$\tfrac{11}{2}\int \tfrac{dx}{(x^2-x+1)^2} = \left|\begin{array}{c} t = x^2-x+1 \\ dt = (2x-1)\,dx \end{array}\right| = \ln|\,x\,| +$$
$$\tfrac{1}{2}\ln\left(x^2-x+1\right) + \tfrac{7}{2}\int \tfrac{dx}{\left(x-\frac{1}{2}\right)^2+\frac{3}{4}} + \tfrac{1}{2}\int \tfrac{dt}{t^2} - \tfrac{11}{2}\int \tfrac{dx}{\left[\left(x-\frac{1}{2}\right)^2+\frac{3}{4}\right]^2} =$$
$$\ln\left|x\sqrt{x^2-x+1}\right| + \tfrac{14}{3}\int \tfrac{dx}{\left(\frac{2x-1}{\sqrt{3}}\right)^2+1} - \tfrac{1}{2t} - \tfrac{88}{9}\int \tfrac{dx}{\left[\left(\frac{2x-1}{\sqrt{3}}\right)^2+1\right]^2} =$$
$$\left|\begin{array}{c} u = \frac{2x-1}{\sqrt{3}} \\ du = \frac{2}{\sqrt{3}}\,dx \end{array}\right| = \ln\left|x\sqrt{x^2-x+1}\right| + \tfrac{7\sqrt{3}}{3}\int \tfrac{du}{u^2+1} - \tfrac{1}{2(x^2-x+1)} -$$
$$\tfrac{44\sqrt{3}}{9}\int \tfrac{du}{\left[u^2+1\right]^2} = \ln\left|x\sqrt{x^2-x+1}\right| + \tfrac{7\sqrt{3}}{3}\operatorname{arctg} u - \tfrac{1}{2(x^2-x+1)} -$$
$$\tfrac{44\sqrt{3}}{9}\left(\tfrac{1}{2}\operatorname{arctg} u + \tfrac{1}{2}\tfrac{u}{u^2+1}\right) + C = \ln\left|x\sqrt{x^2-x+1}\right| +$$
$$\tfrac{7\sqrt{3}}{3}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} - \tfrac{22\sqrt{3}}{9}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} - \tfrac{1}{2(x^2-x+1)} - \tfrac{22\sqrt{3}}{9}\tfrac{\frac{2x-1}{\sqrt{3}}}{\left(\frac{2x-1}{\sqrt{3}}\right)^2+1} + C =$$
$$\ln\left|x\sqrt{x^2-x+1}\right| - \tfrac{\sqrt{3}}{9}\operatorname{arctg}\tfrac{2x-1}{\sqrt{3}} - \tfrac{1}{3}\tfrac{11x-4}{x^2-x+1} + C.$$

□

6.59. Integrate

(a) $\int \frac{x}{1+x^4}\,dx,\ x \in \mathbb{R};$
(b) $\int \frac{5\ln x}{x\ln^3 x + x\ln^2 x - 2x}\,dx,\ x > 0, x \neq \mathrm{e}.$

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,

$$\int \tfrac{x}{1+x^4}\,dx = \left|\begin{array}{c} y = x^2 \\ dy = 2x\,dx \end{array}\right| = \int \tfrac{dy}{2(1+y^2)} = \tfrac{1}{2}\int \tfrac{dy}{1+y^2} =$$
$$\tfrac{1}{2}\operatorname{arctg} y + C = \tfrac{1}{2}\operatorname{arctg} x^2 + C.$$

Case (b). Using substitution, we obtain an integral of a rational function

$$\int \tfrac{5\ln x}{x\ln^3 x + x\ln^2 x - 2x} = \int \tfrac{5\ln x}{\ln^3 x + \ln^2 x - 2}\cdot\tfrac{1}{x}\,dx = \left|\begin{array}{c} y = \ln x \\ dy = \frac{1}{x}\,dx \end{array}\right| =$$
$$\int \tfrac{5y}{y^3+y^2-2}\,dy = \int \tfrac{1}{y-1} + \tfrac{-y+2}{y^2+2y+2}\,dy =$$

Hence it can be immediately seen that this even function will have an absolute maximum at the point $x = 0$ and with the increasing absolute value of x, it will ripple with ever decreasing amplitude.

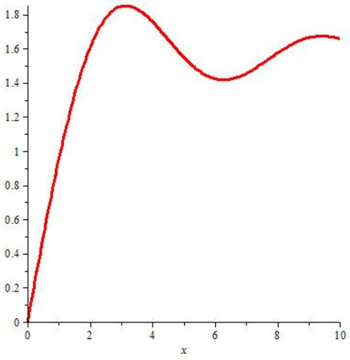
The sine integral function is defined by

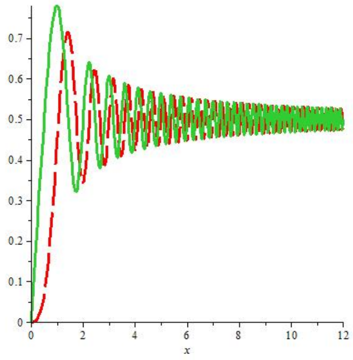
$$\operatorname{Si}(x) = \int_0^x \operatorname{sinc}(t)\,dt.$$

Another important functions are Fresnel's sine and cosine integrals

$$\operatorname{FresnelS}(x) = \int_0^x \sin\left(\tfrac{1}{2}\pi t^2\right)dt$$
$$\operatorname{FresnelC}(x) = \int_0^x \cos\left(\tfrac{1}{2}\pi t^2\right)dt.$$

On the left figure, there's the course of the function $Si(x)$, on the right we can see both Fresnel's functions.





We can also obtain a new type of functions, if we allow a free parameter in the integrated expression, on which the result then depends. As an example, let's look at one of the most important mathematical functions ever — the so calld Gamma function. It's defined by

$$\Gamma(z) = \int_0^\infty \mathrm{e}^{-t}\,t^{z-1}\,dt.$$

It can be shown that this function is analytic at all points $z \notin \mathbb{Z}$ and for small $z \in \mathbb{N}$ we can evaluate:

$$\Gamma(1) = \int_0^\infty \mathrm{e}^{-t}\,t^0\,dt = [-\mathrm{e}^{-t}]_0^\infty = 1$$
$$\Gamma(2) = \int_0^\infty \mathrm{e}^{-t}\,t^1\,dt = [-\mathrm{e}^{-t}\,t]_0^\infty + \int_0^\infty \mathrm{e}^{-t}\,dt = 0 + 1 = 1$$
$$\Gamma(3) = \int_0^\infty \mathrm{e}^{-t}\,t^2\,dt = 0 + 2\int_0^\infty \mathrm{e}^{-t}\,tdt = 0 + 2 = 2$$

and by induction we can easily derive that for all positive integers n this function yields the value of a factorial:

$$\Gamma(n) = (n-1)!$$

The following figure shows the course of the function $f(x) = \ln(\Gamma(x))$ in logarithmic scale of the dependant variable. Hence we can see how fast the factorial actually grows.

$$\int \frac{1}{y-1}\,dy - \frac{1}{2}\int \frac{2y+2}{y^2+2y+2}\,dy + 3\int \frac{1}{(y+1)^2+1^2}\,dy =$$
$$\ln|\,y-1\,| - \frac{1}{2}\ln\left(y^2+2y+2\right) + 3\,\mathrm{arctg}\,(y+1) + C =$$
$$\ln|\,\ln x - 1\,| - \frac{1}{2}\ln\left(\ln^2 x + 2\ln x + 2\right) + 3\,\mathrm{arctg}\,(\ln x + 1) + C.$$

□

For an arbitrary function f that is continuous and bounded on a bounded interval (a,b), the so called Newton-Leibniz formula

$$\int\limits_a^b f(x)\,dx = [F(x)]_a^b := \lim_{x\to b-} F(x) - \lim_{x\to a+} F(x) \qquad (6.9)$$

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 (‖6.9‖). 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).

6.60. 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) $\int \frac{1}{x}\sqrt{\frac{x+1}{x-1}}\,dx,\ x \in \mathbb{R}\setminus[-1,1];$
(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

$$\int f\left(\sqrt[p(1)]{x},\ \sqrt[p(2)]{x},\ldots,\ \sqrt[p(j)]{x}\right)\,dx$$

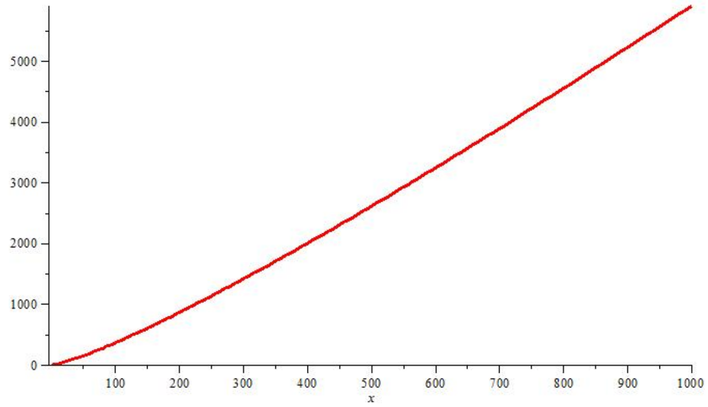
for certain numbers $p(1), p(2), \ldots, p(j) \in \mathbb{N}$ and a rational function f (of more variables), the substitution $t^n = x$ is suggested, where n is the (least) common multiple of numbers $p(1),\ldots,p(j)$. Using this substitution, we can always reduce the integrand (the integrated function) to a rational function, which we can always integrate. We'll get

$$\int \frac{dx}{\sqrt{x^3}+\sqrt[5]{x^7}} = \int \frac{dx}{x\left(\sqrt{x}+\sqrt[5]{x^2}\right)} = \left|\begin{array}{c} t^{10} = x,\ \ \sqrt[10]{x} = t \\ 10t^9\,dt = dx \end{array}\right| = \int \frac{10t^9}{t^{10}\left(t^5+t^4\right)}\,dt =$$
$$10\int \frac{dt}{t^6+t^5} = 10\int \left(\frac{1}{t} - \frac{1}{t^2} + \frac{1}{t^3} - \frac{1}{t^4} + \frac{1}{t^5} - \frac{1}{t+1}\right)dt =$$
$$10\left[\ln t + \frac{1}{t} - \frac{1}{2t^2} + \frac{1}{3t^3} - \frac{1}{4t^4} - \ln(1+t)\right] + C =$$
$$\ln\frac{x}{(1+\sqrt[10]{x})^{10}} + \frac{10}{\sqrt[10]{x}} - \frac{5}{\sqrt[5]{x}} + \frac{10}{3\sqrt[10]{x^3}} - \frac{5}{2\sqrt[5]{x^2}} + C.$$

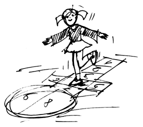
Case (b). For integrals

$$\int f\left(x,\ \sqrt[p(1)]{ax+b},\ \sqrt[p(2)]{ax+b},\ldots,\ \sqrt[p(j)]{ax+b}\right)\,dx,$$

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



Before we plunge into another topics of the mathematical analysis, we'll introduce several more direct applications of the Riemann integral.



6.32. Riemann measurable sets. The definition of the Riemann integral was derived from the concept of size of an area in plane with coordinates x and y bounded the x axis, values of the function $y = f(x)$ and boundary lines $x=a$, $x=b$. Moreover, the area above the x axis is given with a positive sign, while the values under the axis lead to a negative sign. In fact, so far we only know what's an area of a parallelogram determined by two vectors, more generally in vector space $\mathbb{R}^n$ we know what's an area of a parallelepiped. The areas of other subsets are yet to be defined. For some simple objects like polygons, the definition is given naturally by assumed properties.

The concept of the Riemann integral that we bult can now be directly used to measure the "volume" of one-dimensional subsets.

We say 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, i.e. the integral

$$m(A) = \int_{-\infty}^{\infty} \chi_A(x)\,dx$$

exists (the finiteness of its value doesn't matter). The function $\chi_A$is called the *characteristic function of the set* A, the value $m(A)$ is called the *Riemann measure of the set* A. Notice that for an interval $A = [a,b]$ it's actually the value

$$\int_{\infty}^{\infty} \chi_A(x)\,dx = \int_a^b dx = b-a,$$

just as we've expected.

This definition of "size" also has the expected property that the measure of a union of finitely many Riemann measurable pairwise disjoint subsets is the sum of their measures. In particular, every finite set A has zero Riemann measure.

If we instead take a countable union though, this property is no longer true. For example, it suffices to take the set $\mathbb{Q}$ of all rational numbers as a 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.

Notice that the upper Riemann integral of the characteristic set χ_A corresponds to the infimum of the sums of lengths of finitely

$$\int \tfrac{x+1}{\sqrt[3]{3x+1}}\,dx = \left|\begin{array}{c} t^3 = 3x+1 \\ dx = t^2\,dt \end{array}\right| = \int \tfrac{\frac{t^3-1}{3}+1}{t}\,t^2\,dt = \int \tfrac{t^3-1+3}{3}\,t\,dt =$$
$$\tfrac{1}{3}\int t^4+2t\,dt = \tfrac{1}{3}\left(\tfrac{t^5}{5}+t^2\right)+C = \tfrac{t^2}{3}\left(\tfrac{t^3}{5}+1\right)+C =$$
$$\tfrac{\sqrt[3]{(3x+1)^2}}{3}\left(\tfrac{3x+1}{5}+1\right)+C = \sqrt[3]{(3x+1)^2}\,\tfrac{x+2}{5}+C.$$

Case(c). Another generalizations are the integrals of the type

$$\int f\left(x,\ \sqrt[p(1)]{\tfrac{ax+b}{cx+d}},\ \sqrt[p(2)]{\tfrac{ax+b}{cx+d}},\dots,\ \sqrt[p(j)]{\tfrac{ax+b}{cx+d}}\right)\,dx,$$

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,

$$\int \tfrac{1}{x}\sqrt{\tfrac{x+1}{x-1}}\,dx = \left|\begin{array}{c} t^2 = \frac{x+1}{x-1} \\ x = \frac{t^2+1}{t^2-1} \\ dx = -\frac{4t}{(t^2-1)^2}\,dt \end{array}\right| = \int \tfrac{t^2-1}{t^2+1}\,\tfrac{-4t^2}{(t^2-1)^2}\,dt =$$
$$\int \tfrac{-4t^2}{(t^2+1)(t^2-1)}\,dt = \int\left(\tfrac{1}{t+1}-\tfrac{1}{t-1}-\tfrac{2}{t^2+1}\right)dt =$$
$$\ln|\,t+1\,| - \ln|\,t-1\,| - 2\operatorname{arctg} t + C =$$
$$\ln\left|\sqrt{\tfrac{x+1}{x-1}}+1\right| - \ln\left|\sqrt{\tfrac{x+1}{x-1}}-1\right| - 2\operatorname{arctg}\sqrt{\tfrac{x+1}{x-1}}+C.$$

The simplifications

$$\ln\left|\sqrt{\tfrac{x+1}{x-1}}+1\right| - \ln\left|\sqrt{\tfrac{x+1}{x-1}}-1\right| = \ln\left|\frac{\sqrt{\frac{x+1}{x-1}}+1}{\sqrt{\frac{x+1}{x-1}}-1}\right| = \ln\left|\frac{\sqrt{\frac{|\,x+1\,|}{|\,x-1\,|}}+1}{\sqrt{\frac{|\,x+1\,|}{|\,x-1\,|}}-1}\right| =$$
$$\ln\left|\tfrac{\sqrt{|\,x+1\,|}+\sqrt{|\,x-1\,|}}{\sqrt{|\,x+1\,|}-\sqrt{|\,x-1\,|}}\right| = \ln\tfrac{\left(\sqrt{|\,x+1\,|}+\sqrt{|\,x-1\,|}\right)^2}{|\,|\,x+1\,|-|\,x-1\,|\,|} =$$
$$2\ln\left(\sqrt{|\,x+1\,|}+\sqrt{|\,x-1\,|}\right) - \ln 2$$

for $x \in (-\infty, -1) \cup (1, \infty)$ then allow to write

$$\int \tfrac{1}{x}\sqrt{\tfrac{x+1}{x-1}}\,dx = 2\ln\left(\sqrt{|\,x+1\,|}+\sqrt{|\,x-1\,|}\right) - 2\operatorname{arctg}\sqrt{\tfrac{x+1}{x-1}}+C.$$

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

$$\sqrt{ax^2+bx+c} = \sqrt{a}\sqrt{(x-x_1)^2\tfrac{x-x_2}{x-x_1}} = \sqrt{a}\ |\,x-x_1\,|\ \sqrt{\tfrac{x-x_2}{x-x_1}}$$

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)

many disjoint intervals, by which we can cover the given set A, while 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 can proceed in the same way in higher dimensions when defining the *Jordan measure*. For the definition of area (volume) in higher-dimensional space we will also be able to use the concept of the Riemann integral as well, when we generalize it for the multidimensional case. It's good to already notice though that the original notion about an area of a plane figure bounded by a graph of a function in the way described above will indeed be fulfilled completely.

6.33. Mean value of a function. For a finite set of values, we're used to think of their mean value and usually define it as the arithmetic mean.

For a Riemann integrable function $f(x)$ on an interval (finite or infinite) $[a, b]$, its *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 the area between the x axis and the graph of the $f(x)$. Hence the *integral mean value theorem* holds in general.

Proposition. *If $f(x)$ is a Riemann integrable function on an interval $[a, b]$, then there exists a number $m(f)$ satisfying*

$$\int_a^b f(x)\,dx = m(f)(b-a).$$

6.34. Length of a space curve. The integral we built can be also effectively used to compute the *length of a curve* in multidimensional vector space $\mathbb{R}^n$. For the sake of simplicity, we'll show this on the case of a curve in $\mathbb{R}^2$ with coordinates x, y. Suppose we have a parametric description of a curve $F : \mathbb{R} \to \mathbb{R}^2$,

$$F(t) = [g(t), f(t)]$$

and look at it as a trajectory of a movement. For simplicity suppose that $f(t)$ and $g(t)$ have sequentially continuous derivatives.

By differentiating the map $F(t)$ we'll obtain values that will correspond 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 $t = a$, $t = b$) will be given by an integral over the interval $[a, b]$, where the integrated function $h(t)$ will be exactly the sizes of the vectors $F'(t)$. Therefore we want to compute the length s given by

$$s = \int_a^b h(t)\,dt = \int_a^b \sqrt{(f'(t))^2+(g'(t))^2}\,dt.$$

In a special case when the curve is a graph of a function $y = f(x)$ between points $a < b$, we'll obtain

$$s = \int_a^b \sqrt{1+(f'(x))^2}\,dx.$$

The same result can be intuitively seen as a corollary of Pythagor's theorem: for a linear increment of the length of a curve Δs corresponding to the increment Δx of variable x, we can compute

$$\Delta s = \sqrt{(\Delta x)^2+(\Delta y)^2}\,,$$

$$\int \frac{dx}{(x+4)\sqrt{x^2+3x-4}} = \int \frac{dx}{(x+4)\sqrt{(x-1)(x+4)}} = \int \frac{dx}{(x+4)|\,x+4\,|\sqrt{\frac{x-1}{x+4}}} =$$

$$\left| \begin{array}{rl} t^2 &= \frac{x-1}{x+4} \\ x &= \frac{5}{1-t^2} - 4 \\ dx &= \frac{10t}{(1-t^2)^2}\, dt \end{array} \right| = \int \frac{\frac{10t}{(1-t^2)^2}}{\left(\frac{5}{1-t^2}\right)\left|\,\frac{5}{1-t^2}\,\right| t}\, dt = \int \frac{2}{5}\frac{|\,1-t^2\,|}{1-t^2}\, dt =$$

$$\frac{2}{5}\operatorname{sgn}\left(1-t^2\right)\int 1\, dt = \frac{2}{5}\operatorname{sgn}\left(\frac{5}{x+4}\right) t + C =$$

$$\frac{2}{5}\operatorname{sgn}(x)\sqrt{\frac{x-1}{x+4}} + C;$$

(e)

$$\int \frac{dx}{1+\sqrt{-x^2+x+2}} = \int \frac{dx}{1+\sqrt{-(x-2)(x+1)}} = \int \frac{dx}{1+(x+1)\sqrt{\frac{2-x}{x+1}}} =$$

$$\left| \begin{array}{rl} t^2 &= \frac{2-x}{x+1} \\ x &= \frac{3}{t^2+1} - 1 \\ dx &= \frac{-6t}{(t^2+1)^2}\, dt \end{array} \right| = \int \frac{\frac{-6t}{(t^2+1)^2}}{1+\frac{3}{t^2+1}\, t}\, dt = \int \frac{-6t}{(t^2+1)^2}\,\frac{t^2+1}{t^2+3t+1}\, dt =$$

$$\int \frac{-6t}{(t^2+1)(t^2+3t+1)}\, dt =$$

$$\int \left(-\frac{4}{5}\frac{\sqrt{5}}{2t+3+\sqrt{5}} - \frac{2}{t^2+1} - \frac{4}{5}\frac{\sqrt{5}}{-2t-3+\sqrt{5}}\right) dt =$$

$$-\frac{2\sqrt{5}}{5}\ln\left|2t+3+\sqrt{5}\right| - 2\operatorname{arctg} t +$$

$$\frac{2\sqrt{5}}{5}\ln\left|-2t-3+\sqrt{5}\right| + C =$$

$$-\frac{2\sqrt{5}}{5}\ln\left|2\sqrt{\frac{2-x}{x+1}}+3+\sqrt{5}\right| - 2\operatorname{arctg}\sqrt{\frac{2-x}{x+1}} +$$

$$\frac{2\sqrt{5}}{5}\ln\left|-2\sqrt{\frac{2-x}{x+1}}-3+\sqrt{5}\right| + C =$$

$$\frac{2\sqrt{5}}{5}\ln\frac{2\sqrt{\frac{2-x}{x+1}}+3-\sqrt{5}}{2\sqrt{\frac{2-x}{x+1}}+3+\sqrt{5}} - 2\operatorname{arctg}\sqrt{\frac{2-x}{x+1}} + C;$$

(f)

$$\int \frac{dx}{(x-1)\sqrt{x^2+x+1}} = \left| \begin{array}{c} \sqrt{x^2+x+1} = x+t \\ x^2+x+1 = x^2+2xt+t^2 \\ x = -\frac{t^2+2t-2}{2t-1} + 1 \\ dx = \frac{-2(t^2-t+1)}{(2t-1)^2}\, dt \end{array} \right| =$$

$$\int \frac{\frac{-2(t^2-t+1)}{(2t-1)^2}}{-\frac{t^2+2t-2}{2t-1}\,\frac{t^2-t+1}{2t-1}}\, dt = \int \frac{2}{t^2+2t-2}\, dt =$$

$$\int \left(\frac{\sqrt{3}}{3}\frac{1}{t+1-\sqrt{3}} - \frac{\sqrt{3}}{3}\frac{1}{t+1+\sqrt{3}}\right) dt =$$

$$\frac{\sqrt{3}}{3}\ln\left|t+1-\sqrt{3}\right| - \frac{\sqrt{3}}{3}\ln\left|t+1+\sqrt{3}\right| + C =$$

$$\frac{\sqrt{3}}{3}\ln\left|\,\frac{t+1-\sqrt{3}}{t+1+\sqrt{3}}\,\right| + C = \frac{\sqrt{3}}{3}\ln\frac{\left|\sqrt{x^2+x+1}-x+1-\sqrt{3}\right|}{\sqrt{x^2+x+1}-x+1+\sqrt{3}} + C.$$

□

6.61. 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

and when looking at our definition of integral, that means

$$s = \int_a^b \sqrt{1+\left(\frac{dy}{dx}\right)^2}\, dx.$$

Conversely, the fundamental theorem of differential calculus (see 6.25) shows, that at the level of differentials, such defined quantity of the length of a graph of a function $y = y(x)$ satisfies

$$ds = \sqrt{1+(y'(x))^2}\, dx,$$

just as we've expected.

As an easy example, we'll calculate the length of a unit circle as a double of an integral of the function $y = \sqrt{1-x^2}$ over $[-1, 1]$. We already know that the result must be 2π, because we defined π in this way.

$$\begin{aligned} s &= 2\int_{-1}^{1}\sqrt{1+(y')^2}\, dx = 2\int_{-1}^{1}\sqrt{1+\frac{x^2}{1-x^2}}\, dx \\ &= 2\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}\, dx = 2[\arcsin x]_{-1}^{1} = 2\pi. \end{aligned}$$

If we instead use

$$y = \sqrt{r^2-x^2} = r\sqrt{1-(x/r)^2}$$

and bounds $[-r, r]$ in the previous calculation, by substituting $x = rt$ we'll obtain the length of a circle with radius r:

$$\begin{aligned} s(r) &= 2\int_{-r}^{r}\sqrt{1+\frac{(x/r)^2}{1-(x/r)^2}}\, dx = 2\int_{-1}^{1}\frac{r}{\sqrt{1-t^2}}\, dt \\ &= 2r[\arcsin x]_{-1}^{1} = 2\pi r. \end{aligned}$$

The result is of course well known from elementary geometry. Nonetheless, by using integral calculus, we've just derived an important fact, that the length of a circle is linearly dependent on its diameter $2r$. The number π is exactly the ration, in which is this dependancy realized.

6.35. Areas and volumes. The Riemann integral can be used directly to compute areas or volumes of shapes defined by a graph of a function.

As an example, let's calculate the area of a circle with radius r. The halfcircle bounded by the function $\sqrt{r^2-x^2}$ has an area, whose double $a(r)$ can be computed using the substitution $x = r\sin t$, $dx = r\cos t\, dt$ (using the corollary for I_2 in the paragraph 6.22)

$$\begin{aligned} a(r) &= 2\int_{-r}^{r}\sqrt{r^2-x^2}\, dx = 2r^2\int_{-\pi/2}^{\pi/2}\cos^2 t\, dt \\ &= \frac{2r^2}{2}[\cos t\sin t + t]_{-\pi/2}^{\pi/2} = \pi r^2. \end{aligned}$$

It's again worth noticing that this well known formula is derived from the principles of integral calculus and surprisingly, 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.

Also notice the ratio of the area and the perimeter of a circle, i.e.

$$\frac{\pi r^2}{2\pi r} = \frac{r}{2}.$$

A square with the same area has a side of length $\sqrt{\pi}r$ and therefore its perimeter is $4\sqrt{\pi}r$. Hence the perimeter of a square with an

$$\int \tfrac{dx}{x+\sqrt{x^2+x-1}} = \left| \begin{array}{c} \sqrt{x^2+x-1} = x+t \\ x^2+x-1 = x^2+2xt+t^2 \\ x = \frac{t^2+1}{1-2t} \\ dx = \frac{-2t^2+2t+2}{(1-2t)^2}\, dt \end{array} \right| = \int \tfrac{-2t^2+2t+2}{(t+2)(1-2t)}\, dt =$$

$$\int \left(1 - \tfrac{2}{t+2} - \tfrac{1}{2}\tfrac{1}{t-\frac{1}{2}}\right) dt = t - 2\ln|t+2| - \tfrac{1}{2}\ln\left|t-\tfrac{1}{2}\right| + C =$$

$$\sqrt{x^2+x-1} - x - 2\ln\left(\sqrt{x^2+x-1} - x + 2\right) -$$

$$\tfrac{1}{2}\ln\left|\sqrt{x^2+x-1} - x - \tfrac{1}{2}\right| + C.$$

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. □

Another method of integration can be found on page 406

D. Definite integrals

6.62. Compute the definite integrals

$$\int\limits_{\frac{\pi}{6}}^{\frac{\pi}{3}} \mathrm{tg}^2\, x\ dx, \qquad \int\limits_{0}^{\frac{\pi}{4}} \tfrac{x}{\cos^2 x}\ dx.$$

Solution. For $x \neq \frac{\pi}{2} + k\pi$, where $k \in \mathbb{Z}$, we have

$$\int \mathrm{tg}^2\, x\, dx = \mathrm{tg}\, x - x + C,$$

as we have compute earlier. This implies that

$$\int\limits_{\pi/6}^{\pi/3} \mathrm{tg}^2\, x\, dx = \left[\mathrm{tg}\, x - x\right]_{\pi/6}^{\pi/3} = \sqrt{3} - \tfrac{\pi}{3} - \left(\tfrac{1}{\sqrt{3}} - \tfrac{\pi}{6}\right) = \tfrac{2}{\sqrt{3}} - \tfrac{\pi}{6}.$$

Of course, definite integrals can be also computed directly. For example, the substitution $y = \mathrm{tg}\, x$ yields

$$\int\limits_{\pi/6}^{\pi/3} \mathrm{tg}^2\, x\, dx = \int\limits_{\pi/6}^{\pi/3} \tfrac{\sin^2 x}{\cos^2 x}\, dx = \left| \begin{array}{c} y = \mathrm{tg}\, x;\, dy = \frac{dx}{\cos^2 x} \\ \sin^2 x = \frac{\mathrm{tg}^2\, x}{1+\mathrm{tg}^2\, x} = \frac{y^2}{1+y^2} \end{array} \right| =$$

$$\int\limits_{1/\sqrt{3}}^{\sqrt{3}} \tfrac{y^2}{1+y^2}\, dy = \int\limits_{1/\sqrt{3}}^{\sqrt{3}} 1 - \tfrac{1}{1+y^2}\, dy = \left[y - \mathrm{arctg}\, y\right]_{1/\sqrt{3}}^{\sqrt{3}} = \tfrac{2}{\sqrt{3}} - \tfrac{\pi}{6}.$$

When doing the substitution, we only need to not forget to change the limits of the integral to values gained by substituting $\sqrt{3} = \mathrm{tg}\,(\pi/3)$, $1/\sqrt{3} = \mathrm{tg}\,(\pi/6)$.

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

$$\int\limits_{0}^{\pi/4} \tfrac{x}{\cos^2 x}\, dx = \left| \begin{array}{c|c} F(x) = x & F'(x) = 1 \\ G'(x) = \frac{1}{\cos^2 x} & G(x) = \mathrm{tg}\, x \end{array} \right| =$$

$$\left[x\, \mathrm{tg}\, x\right]_0^{\pi/4} - \int\limits_0^{\pi/4} \mathrm{tg}\, x\, dx = \left[x\, \mathrm{tg}\, x\right]_0^{\pi/4} + \int\limits_0^{\pi/4} \tfrac{-\sin x}{\cos x}\, dx =$$

$$\left[x\, \mathrm{tg}\, x\right]_0^{\pi/4} + [\ln(\cos x)]_0^{\pi/4} = \tfrac{\pi}{4} + \ln\tfrac{\sqrt{2}}{2} = \tfrac{\pi - 2\ln 2}{4}.$$

□

area of a unit circle is $4\sqrt{\pi}$, which is approximately 0.8 more than the perimeter of a unit circle. It can be shown, in fact, that circle is a shape with the smallest perimeter among all with the same area. We'll get to derive such results in our comments about the so called calculus of variations in later chapters.

Another analogy of the this principle is the computation of *surface and volume of a solid of revolution.* If a solid originates by the rotation of a graph of function f around the x axis in an interval $[a, b]$, an increment Δx causes the area to increase by a multiple of the length Δs of the curve given by the graph a of function $y = f(x)$ and the size of a circle with radius $f(x)$. Hence the surface can be 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 ds is given by the increment on the length of curve $y = f(x)$, see above. If we would determine the solid of revolution by its bound parametrized by a pair of functions $[x(t), y(t)]$, the corresponding differential will be in the form $ds = \sqrt{(x'(t))^2 + (y'(t))^2}\, dt$ and for the surface, we'll get

$$A = 2\pi \int_a^b y(t)\sqrt{(y'(t))^2 + (x'(t))^2}\ dt.$$

When changing Δx, the volume of the same solid will increase by a multiple of this increment and the area of a circle with radius $f(x)$. Hence it's given by the formula

$$V(f) = \pi \int_a^b (f(x))^2\, dx.$$

As an example of using the formulas for surface and volume, we'll derive the well known formulas for the surface of a sphere and volume of a ball with diameter r.

$$\begin{aligned} A_r &= 2\pi \int_{-r}^{r} r\sqrt{1 - (x/r)^2}\, \frac{1}{\sqrt{1 - (x/r)^2}}\, dt \\ &= 2\pi r \int_{-r}^{r} dt = 4\pi r^2 \\ V_r &= \pi \int_{-r}^{r} (r^2 - x^2)\, dx \\ &= \pi \left[r^2 x - \frac{1}{3}x^3\right]_{-r}^{r} = \frac{4}{3}\pi r^3. \end{aligned}$$

Similar to the circle, a ball is also and object which has the smallest volume among all with a given surface. That's the reason why soap bubbles almost always assume this shape.

6.36. Integral criterion of the convergence of series. Using the improper integral, we can also decide the question of convergence of a wider class of infinite series than before:

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 we interpret the integral as an area under the curve, the criterion is clear.

6.63. 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) $\int_0^1 \left(\frac{\mathrm{e}^x}{\mathrm{e}^{2x}+3} + \frac{1}{\cos^2 x}\right) dx$;

Solution. We have

(a)

$$\int\limits_0^1 \frac{x}{\sqrt{1-x^2}}\, dx = \left| \begin{array}{c} y = 1-x^2 \\ dy = -2x\, dx \end{array} \right| = -\int\limits_1^0 \frac{y^{-1/2}}{2}\, dy =$$

$$\int\limits_0^1 \frac{y^{-1/2}}{2}\, dy = \left[\sqrt{y}\,\right]_0^1 = 1;$$

(b)

$$\int\limits_1^2 \frac{dx}{\sqrt{x^2-1}} = \left| \begin{array}{c} z = x + \sqrt{x^2-1} \\ dz = \frac{\sqrt{x^2-1}+x}{\sqrt{x^2-1}}\, dx \end{array} \right| = \int\limits_1^{2+\sqrt{3}} \frac{1}{z}\, dz =$$

$$[\ln z]_1^{2+\sqrt{3}} = \ln\left(2+\sqrt{3}\right);$$

(c)

$$\int\limits_0^1 \left(\frac{\mathrm{e}^x}{\mathrm{e}^{2x}+3} + \frac{1}{\cos^2 x}\right) dx = \int\limits_0^1 \frac{\mathrm{e}^x}{\mathrm{e}^{2x}+3}\, dx + \int\limits_0^1 \frac{1}{\cos^2 x}\, dx =$$

$$\left| \begin{array}{c} p = \mathrm{e}^x \\ dp = \mathrm{e}^x\, dx \end{array} \right| = \int\limits_1^{\mathrm{e}} \frac{1}{p^2+3}\, dp + [\mathrm{tg}\, x]_0^1 = \frac{1}{3}\int\limits_1^{\mathrm{e}} \frac{1}{\left(\frac{p}{\sqrt{3}}\right)^2+1}\, dp +$$

$$\mathrm{tg}\, 1 = \left| \begin{array}{c} s = \frac{p}{\sqrt{3}} \\ ds = \frac{1}{\sqrt{3}}\, dp \end{array} \right| = \frac{\sqrt{3}}{3}\int\limits_{1/\sqrt{3}}^{\mathrm{e}/\sqrt{3}} \frac{1}{s^2+1}\, ds + \mathrm{tg}\, 1 =$$

$$\frac{\sqrt{3}}{3}[\mathrm{arctg}\, s]_{1/\sqrt{3}}^{\mathrm{e}/\sqrt{3}} + \mathrm{tg}\, 1 = \frac{\sqrt{3}}{3}\left(\mathrm{arctg}\, \frac{\mathrm{e}\sqrt{3}}{3} - \frac{\pi}{6}\right) + \mathrm{tg}\, 1;$$

□

6.64. Prove that

$$\frac{\sqrt{2}}{20} \le \int\limits_0^1 \frac{x^9}{\sqrt{1+x}}\, dx \le \frac{1}{10}.$$

Solution. Because

$$0 \le \frac{x^9}{\sqrt{2}} \le \frac{x^9}{\sqrt{1+x}} \le x^9, \quad x \in [0,1],$$

the geometric meaning of the definite integral implies

$$\frac{\sqrt{2}}{20} = \int\limits_0^1 \frac{x^9}{\sqrt{2}}\, dx \le \int\limits_0^1 \frac{x^9}{\sqrt{1+x}}\, dx \le \int\limits_0^1 x^9\, dx = \frac{1}{10}.$$

□

6.65. Without symbols of differentiation and integration, express

$$\left(\int\limits_x^0 t^5 \ln(t+1)\, dt\right)', \quad x \in (-1,1),$$

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

$$F(x) := \int\limits_0^x t^5 \ln(t+1)\, dt, \quad x \in (-1,1)$$

If the given series diverges, then the series $\sum_{n=2}^{\infty} f(n)$ diverges as well. For any $k \in \mathbb{N}$, we have the following inequality for k-th partial sum s'_k (of the series without its first term)

$$s'_k = \sum_{n=2}^{k} f(n) < \int_1^k f(x)\,\mathrm{d}x,$$

because s'_k is a lower sum of the Riemann integral $\int_1^k f(x)\,\mathrm{d}x$. But then

$$\int_1^\infty f(x)\,\mathrm{d}x = \lim_{k\to\infty}\int_1^k f(x)\,\mathrm{d}x > \lim_{k\to\infty} s'_k = \infty$$

and the intergral diverges.

Now suppose the given integral converges and denote the k-th partial sum of the given series by s_k. Then we have the inequalities

$$\int_1^\infty f(x)\,\mathrm{d}x = \lim_{k\to\infty}\int_1^k f(x)\,\mathrm{d}x < \lim_{k\to\infty} s_k < \infty,$$

because s_k is an upper sum of the Riemann integral $\int_1^k f(x)\,\mathrm{d}x$ and we suppose the given series converges. □

3. Infinite series

While building our menagerie, we have already encountered power series, which extend the collection of all polynomials in a natural way, see 5.44. We also said that we'll obtain a class of analytic functions in this way, but we didn't even prove that power series are continuous functions. Now we can easily show that it is indeed the case and that we can also differentiate and integrate power series term by term. Because of this, we will see that it's not possible to obtain a sufficiently wide class of functions by using power series. For example, in this way we can never obtain only sequentially continuous periodic functions, which are very important for simulations and processing of audio and video signals.

6.37. How tamed are our series of functions? Let's now return

to disscussing the limits of sequences of functions and the sum of series of functions from the point of applying 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 are:

- 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?

First we'll show on examples that the answer on all three such formulated questions are "NO!". But then we'll find simple additional conditions on the convergence of the series which, on the contrary, will guarantee the validity of all three statements. Hence the series of functions are not generally well managable, though we can choose a wide class of ones which can be worked with very is clearly the antiderivative of function $f(x) := x^5 \ln(x+1)$ on interval $(-1, 1)$, i.e. by differentiating it, we'll get exactly f. Hence

$$\left(\int_x^0 t^5 \ln(t+1)\, dt\right)' = -\left(\int_0^x t^5 \ln(t+1)\, dt\right)' = -x^5 \ln(x+1)\,.$$

□

E. Improper integrals

6.66. Decide if

$$\int_1^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\, dx \in \mathbb{R}.$$

Solution. The improper integral represents the area of the figure between the graph of a positive function

$$y = \frac{\operatorname{arctg} x}{x\sqrt{x}}, \qquad x \geq 1$$

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

$$\tfrac{\pi}{4} \leq \operatorname{arctg} x \leq \tfrac{\pi}{2}, \qquad x \in [1, +\infty).$$

But that implies

$$\tfrac{\pi}{2} = \tfrac{\pi}{4} \int_1^{+\infty} x^{-\frac{3}{2}}\, dx \leq \int_1^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\, dx \leq \tfrac{\pi}{2} \int_1^{+\infty} x^{-\frac{3}{2}}\, dx = \pi,$$

i.e. in particular

$$\int_1^{+\infty} \frac{\operatorname{arctg} x}{x\sqrt{x}}\, dx \in \mathbb{R}.$$

□

The formula (‖6.9‖) 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 $a \in \mathbb{R}$).

6.67. Determine

(a) $\int_1^{\infty} \sin x \; dx\,;$

(b) $\int_1^{\infty} \frac{dx}{x^4+x^2}\,;$

(c) $\int_0^{4} \frac{dx}{\sqrt{x}}\,;$

(d) $\int_{-1}^{1} \frac{dx}{x^2}\,.$

Solution. Case (a). We can immediately determine

$$\int_1^{\infty} \sin x\, dx = [-\cos x]_1^{\infty} = \lim_{x\to\infty} (-\cos x) + \cos 1.$$

well. Fortunately, power series will belong there as well. Then we'll also give some thoughts to alternative concepts of integration that work more satisfyingly even for wider classes of functions

6.38. Examples of nasty sequences. (1) First consider the functions

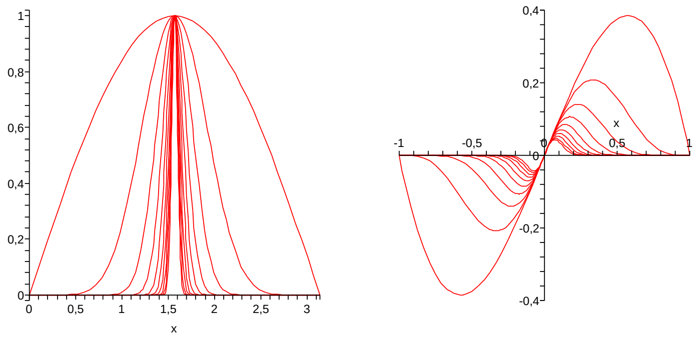
$$f_n(x) = (\sin x)^n$$

on intervalu $[0, \pi]$. The values of these functions will be nonnegative and lesser than one at all points $0 \leq x \leq \pi$, except for $x = \frac{\pi}{2}$, where the value is 1. Hence on the whole interval $[0, \pi]$, these functions will 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. Clearly, the limit of the sequence of functions f_n is a noncontinuous function, even though all functions $f_n(x)$ are continuous. The problematic point is even an inner point of the interval.

We can find the same phenomenon for 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-t partial sum. For example, $f_1(x) = \sin x$, $f_2(x) = (\sin x)^2 - \sin x$, etc. The left figure plots the functions $f_m(x)$ for $m = n^3$, $n = 1, \dots, 10$.



(2) Let's now look at the second question, i.e. badly behaving derivatives. Quite natural idea on the same principle as above is constructing a sequence of functions which will always have the same nonzero derivative at one point, but they will become smaller and smaller, so they will pointwise converge to an identic zero function.

The previous figure on the right plots the functions

$$f_n(x) = x(1-x^2)^n$$

on interval $[-1, 1]$ for values $n = m^2$, $m = 1, \dots, 10$. At first glance, it's clear that

$$\lim_{n\to\infty} f_n(x) = 0$$

and all functions $f_n(x)$ are smooth. Their derivative at the point $x = 0$ is

$$f_n'(0) = \left((1-x^2)^n - 2nx^2(1-x^2)^{n-1}\right)|_{x=0} = 1$$

no matter the n. But the limit function for the sequence f_n has a zero derivative at every point of course!

(3) We've already seen the counterexample to the third statement in 6.32. The characteristic function $\chi_{\mathbb{Q}}$ of rational numbers can be expressed as a sum of countably many functions, which will be numbered exactly by rational numbers and will be zero everywhere except for the single point after which they are named for,

Because the limit on the right hand side doesn't exist, the integral oscilates.

Cases (b), (c). Analogously, we can easily compute

$$\int_1^\infty \frac{dx}{x^4+x^2} = \int_1^\infty \frac{dx}{x^2(x^2+1)} = \int_1^\infty \frac{1}{x^2} - \frac{1}{1+x^2}\,dx = \left[-\frac{1}{x} - \operatorname{arctg} x\right]_1^\infty = \lim_{x\to\infty}\left(-\frac{1}{x} - \operatorname{arctg} x\right) + \frac{1}{1} + \operatorname{arctg} 1 = 0 - \frac{\pi}{2} + 1 + \frac{\pi}{4} = 1 - \frac{\pi}{4}$$

and even more easily

$$\int_0^4 \frac{dx}{\sqrt{x}} = \left[2\sqrt{x}\right]_0^4 = 4 - 0 = 4,$$

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

$$\int_{-1}^1 \frac{dx}{x^2} = \left[-\frac{1}{x}\right]_{-1}^1 = -1 - 1 = -2,$$

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 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

$$\int_{-1}^1 \frac{dx}{x^2} = \int_{-1}^0 \frac{dx}{x^2} + \int_0^1 \frac{dx}{x^2} = \left[-\frac{1}{x}\right]_{-1}^0 + \left[-\frac{1}{x}\right]_0^1 = \lim_{x\to 0-}\left(-\frac{1}{x}\right) - 1 - 1 - \lim_{x\to 0+}\left(-\frac{1}{x}\right) = \infty - 2 + \infty = +\infty.$$

Note that the even character of function $y = x^{-2}$ also implies

$$\int_{-1}^1 \frac{dx}{x^2} = 2\int_0^1 \frac{dx}{x^2} = 2\cdot\infty = +\infty.$$

□

6.68. 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)

$$\int_0^\infty \frac{dx}{(x+2)^5} = -\frac{1}{4}\left[(x+2)^{-4}\right]_0^\infty = -\frac{1}{4}\left(\lim_{x\to\infty}(x+2)^{-4} - 2^{-4}\right) = -\frac{1}{4}\left(0 - \frac{1}{16}\right) = \frac{1}{64};$$

(b)

$$\int_{-2}^2 \ln|\,x\,|\,dx = \int_{-2}^0 \ln|\,x\,|\,dx + \int_0^2 \ln|\,x\,|\,dx = 2\int_0^2 \ln x\,dx = \left|\begin{array}{l|l} F(x) = \ln x & F'(x) = \frac{1}{x} \\ G'(x) = 1 & G(x) = x \end{array}\right| = 2\left([x\ln x]_0^2 - \int_0^2 1\,dx\right) =$$

where they value will be 1. Riemann integrals of all such functions will be zero, but their sum is not a Riemann integrable function.

This example illustrates the fundamental flaw of the Riemann integral, which we'll come back to later.

We can easily also find an example when the limit function f is integrable, all functions f_n are continuous, but the value of the integral still isn't the limit of the values of integrals of f_n. It suffices to slightly change the sequence of functions which we used above:

$$f_n(x) = 2nx(1-x^2)^n.$$

We can easily verify that the values of these functions also converge to zero for every $x \in [0,1]$ (for example we can see that $\ln(f_n(x)) \to -\infty$). But

$$\int_0^1 f_n(x)\,dx = \frac{n}{n+1} \to 1 \neq 0.$$

6.39. Uniform convergence. An obvious 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 limit ourselves to cases where the convergence will have roughly the same speed all over the interval

Uniform convergence

Definition. We say the sequence of functions $f_n(x)$ *converges uniformly* on interval $[a,b]$ to a 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.

We say 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's independant on the point $x \in [a,b]$. That's a difference from the pointwise convergence, where N depends on both ε and x. We can 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. Clearly the first and the last of the previous cases didn't have this property; at the second case, the sequence of derivatives f_n' lacked it.

The following three theorems can be briefly summed up by a statement that all three generally false statements in 6.37 are true for uniform convergence (but beware of the subtilities when differentiating).

6.40. Theorem. *Let $f_n(x)$ be a sequence of functions that are continuous on interval $[a,b]$, which converges uniformly to function $f(x)$ on this interval. Then $f(x)$ is also continuous on interval $[a,b]$.*

Proof. We want to show that for an arbitrary fixed point $x_0 \in [a,b]$ and any fixed small $\varepsilon > 0$, the inequality

$$|f(x) - f(x_0)| < \varepsilon$$

$$2\left([x\ln x]_0^2-[x]_0^2\right)=2\left(2\ln 2-\lim_{x\to 0+}(x\ln x)-2+0\right)=$$
$$4\ln 2-4;$$

(c)

$$\int_1^\infty \frac{e^{-\sqrt{x}}}{\sqrt{x}}\,dx=\left|\begin{matrix} t=\sqrt{x}\\ dt=\frac{1}{2\sqrt{x}}\,dx\end{matrix}\right|=2\int_1^\infty e^{-t}\,dt=2\left[-e^{-t}\right]_1^\infty=$$
$$-2\left(\lim_{t\to\infty}e^{-t}-e^{-1}\right)=\frac{2}{e};$$

(d)

$$\int_{-1}^{0}\frac{e^{1/x}}{x^3}\,dx=\left|\begin{matrix} u=1/x\\ du=-\frac{1}{x^2}\,dx\end{matrix}\right|=-\int_{-1}^{-\infty}u\,e^u\,du=$$
$$\int_{-\infty}^{-1}u\,e^u\,du=\left|\begin{matrix} F(u)=u\\ G'(u)=e^u\end{matrix}\right|\begin{matrix} F'(u)=1\\ G(u)=e^u\end{matrix}\right|=$$
$$[u\,e^u]_{-\infty}^{-1}-\int_{-\infty}^{-1}e^u\,du=[u\,e^u]_{-\infty}^{-1}-[e^u]_{-\infty}^{-1}=$$
$$-\frac{1}{e}-\lim_{u\to-\infty}u\,e^u-\frac{1}{e}+\lim_{u\to-\infty}e^u=-\frac{2}{e};$$

(e)

$$\int_1^2\frac{dx}{x\ln x}=\left|\begin{matrix} r=\ln x\\ dr=\frac{1}{x}\,dx\end{matrix}\right|=\int_0^{\ln 2}\frac{dr}{r}=[\ln r]_0^{\ln 2}=$$
$$\ln(\ln 2)-\lim_{r\to 0+}\ln r=\ln(\ln 2)+\infty=+\infty.$$

□

6.69. Compute the improper integrals

$$\int_0^\infty x^2\,e^{-x}\,dx;\qquad \int_{-\infty}^{\infty}\frac{dx}{e^x+e^{-x}}.$$

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

$$\int_0^\infty x^2\,e^{-x}\,dx=\left|\begin{matrix} F(x)=x^2\\ G'(x)=e^{-x}\end{matrix}\right|\begin{matrix} F'(x)=2x\\ G(x)=-e^{-x}\end{matrix}\right|=$$
$$\left[-x^2\,e^{-x}\right]_0^\infty+2\int_0^\infty x\,e^{-x}\,dx=\left|\begin{matrix} F(x)=x\\ G'(x)=e^{-x}\end{matrix}\right|\begin{matrix} F'(x)=1\\ G(x)=-e^{-x}\end{matrix}\right|=$$
$$-\lim_{x\to\infty}\frac{x^2}{e^x}+2\left[-x\,e^{-x}\right]_0^\infty+2\int_0^\infty e^{-x}\,dx=$$
$$0-2\lim_{x\to\infty}\frac{x}{e^x}+2\left[-e^{-x}\right]_0^\infty=0+2\left(\lim_{x\to\infty}-e^{-x}+1\right)=2.$$

The substitution method then yields

$$\int_{-\infty}^{\infty}\frac{dx}{e^x+e^{-x}}=\int_{-\infty}^{\infty}\frac{e^x}{e^{2x}+1}\,dx=\left|\begin{matrix} y=e^x\\ dy=e^x\,dx\end{matrix}\right|=\int_0^\infty\frac{dy}{y^2+1}=$$
$$\left[\operatorname{arctg}y\right]_0^\infty=\lim_{y\to\infty}\operatorname{arctg}y=\frac{\pi}{2},$$

when the new limits of the integral are derived from the limits

$$\lim_{x\to-\infty}e^x=0,\qquad \lim_{x\to\infty}e^x=+\infty.$$

□

6.70. Compute

$$\int_0^\infty x^{2n+1}\,e^{-x^2}\,dx,\qquad n\in\mathbb{N}.$$

will hold for all x close enough to x_0. From the definition of uniform convergence, for some $\varepsilon>0$ we have

$$|f_n(x)-f(x)|<\varepsilon$$

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's possible, because all $f_n(x)$ are continuous). 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 our chosen δ-neighbourhood of x_0. □

6.41. Theorem. *Let $f_n(x)$ be a sequence of Riemann integrable functions on a finite interval $[a,b]$ which converge uniformly to function $f(x)$ on this interval. Then $f(x)$ is Riemann integrable as well and*

$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx=\int_a^b\left(\lim_{n\to\infty}f_n(x)\right)dx=\int_a^b f(x)\,dx.$$

The proof of this theorem is based upon a generalization of properties of Cauchy sequences of numbers to uniform convergence of functions. This way we can work with the existence of the limit of a sequence of integrals without needing to know it.

Uniformly Cauchy sequences

Definition. We say the sequence of functions $f_n(x)$ on interval $[a,b]$ is *uniformly Cauchy*, if for every (small) positive number ε, there exists (large) natural number N such that for all $x\in[a,b]$ and all $n\ge N$, the inequality

$$|f_n(x)-f_m(x)|<\varepsilon$$

holds.

Clearly every uniformly convergent sequence of function on interval $[a,b]$ is also uniformly Cauchy on the same interval; it suffices to notice the usual bound

$$|f_n(x)-f_m(x)|\le|f_n(x)-f(x)|+|f(x)-f_m(x)|$$

based on triangle inequality.

This observation will now suffice to prove our theorem, but first we'll stop at a convinient converse statement:

Proposition. *Every uniformly Cauchy sequence of functions $f_n(x)$ on interval $[a,b]$ uniformly converges to some function f on this interval.*

Proof. The condition for a sequence of functions to be Cauchy implies that also for all $x\in[a,b]$, the sequence of values $f_n(x)$ is a Cauchy sequence of real (eventually complex) numbers. Hence the sequence of functions $f_n(x)$ must converge pointwise to some function $f(x)$.

We'll show that in fact, the sequence $f_n(x)$ converges to its limit uniformly. Choose N large enough so that

$$|f_n(x)-f_m(x)|<\varepsilon$$

Solution. We'll first solve this problem by the substitution method and then repeatedly apply integration by parts, yielding

$$\int\limits_0^\infty x^{2n+1}\,\mathrm{e}^{-x^2}\,dx = \left|\begin{matrix} y = x^2 \\ dy = 2x\,dx \end{matrix}\right| = \tfrac{1}{2}\int\limits_0^\infty y^n\,\mathrm{e}^{-y}\,dy =$$

$$\left|\begin{array}{l|l} F(y) = y^n & F'(y) = ny^{n-1} \\ G'(y) = \mathrm{e}^{-y} & G(y) = -\mathrm{e}^{-y} \end{array}\right| =$$

$$\tfrac{1}{2}\left(\left[-y^n\,\mathrm{e}^{-y}\right]_0^\infty + n\int\limits_0^\infty y^{n-1}\,\mathrm{e}^{-y}\,dy\right) = \tfrac{n}{2}\int\limits_0^\infty y^{n-1}\,\mathrm{e}^{-y}\,dy =$$

$$\left|\begin{array}{l|l} F(y) = y^{n-1} & F'(y) = (n-1)y^{n-2} \\ G'(y) = \mathrm{e}^{-y} & G(y) = -\mathrm{e}^{-y} \end{array}\right| =$$

$$\tfrac{n}{2}\left(\left[-y^{n-1}\,\mathrm{e}^{-y}\right]_0^\infty + (n-1)\int\limits_0^\infty y^{n-2}\,\mathrm{e}^{-y}\,dy\right) =$$

$$\tfrac{n(n-1)}{2}\int\limits_0^\infty y^{n-2}\,\mathrm{e}^{-y}\,dy = \cdots = \tfrac{n(n-1)\cdots 2}{2}\int\limits_0^\infty y\,\mathrm{e}^{-y}\,dy =$$

$$\left|\begin{array}{l|l} F(y) = y & F'(y) = 1 \\ G'(y) = \mathrm{e}^{-y} & G(y) = -\mathrm{e}^{-y} \end{array}\right| = \tfrac{n!}{2}\left(\left[-y\,\mathrm{e}^{-y}\right]_0^\infty + \int\limits_0^\infty \mathrm{e}^{-y}\,dy\right) =$$

$$\tfrac{n!}{2}\left[-\mathrm{e}^{-y}\right]_0^\infty = \tfrac{n!}{2}.$$

□

6.71. In dependancy on $a \in \mathbb{R}^+$ determine the integral $\int_0^1 \frac{1}{x^a}\,\mathrm{d}x$. ○

F. Lengths, areas, surfaces, volumes

6.72. 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

$$\int_0^{\frac{\pi}{2}} \sqrt{(x'(t))^2 + (y'(t))^2}\,\mathrm{d}t = \int_0^{\frac{\pi}{2}} \sqrt{(\sin 2t)^2 + (-\sin 2t)^2}\,\mathrm{d}t$$
$$= \int_0^{\frac{\pi}{2}} \sqrt{2}\sin 2t\,\mathrm{d}t = \sqrt{2}.$$

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 $\sqrt{2}$. □

6.73. 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 ‖**??**‖:

$$l = \int_0^{\sqrt{5}} \sqrt{4t^2 + 9t^4}\,\mathrm{d}t = \int_0^{\sqrt{5}} t\sqrt{9t^2 + 4}\,\mathrm{d}t$$
$$= \frac{1}{2}\int_0^5 \sqrt{9u + 4}\,\mathrm{d}t = \frac{2}{27}[(9u + 4)^{\frac{3}{2}}]_0^5 = \frac{335}{27}$$

□

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 we have

$$|f_n(x) - f(x)| = \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \varepsilon$$

for all $x \in [a, b]$. □

Proof of the Theorem. Recall that every uniformly convergent sequence of functions is also uniformly Cauchy and that the Riemann sums of all single terms of our sequence converge to $\int_a^b f_n(x)\,dx$ independently on the choice of the partition and the representants. Hence, if we have

$$|f_n(x) - f_m(x)| < \varepsilon$$

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, hence convergent. Also because of the uniform convergence of the sequence $f_n(x)$, the same must be true for the limit function $f(x)$ (its Riemann sums are arbitrarily close to the Riemann sums of the functions f_n for sufficiently large n), so the limit function $f(x)$ will again be integrable. Moreover,

$$\left|\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx\right| \leq \varepsilon|b - a|,$$

so it must be the correct limit value. □

For the corresponding result about derivatives, we need to take extra care regarding the assumptions:

6.42. Theorem. *Let $f_n(x)$ be a sequence of functions differentiable on interval $[a, b]$ and assume $f_n(x_0) \to f(x_0)$ at some point $x_0 \in [a, b]$. Moreover, let all derivatives $g_n(x) = f_n'(x)$ be continuous and let them converge uniformly to function $g(x)$ on the same interval. The the function $f(x) = \int_{x_0}^x g(t)\,dt$ is also differentiable on interval $[a, b]$, the functions $f_n(x)$ converge to $f(x)$ and $f'(x) = g(x)$.*

Proof. If we consider functions $\tilde{f}_n(x) = f_n(x) - f_n(x_0)$ instead of $f_n(x)$, the assumptions and conclusions in the theorem will be valid or invalid for both sequences and the same time. Hence without loss of generality we can assume that all our functions satisfy $f_n(x_0) = 0$. Then 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 uniformly converge to function g on whole $[a, b]$, the functions $f_n(x)$ converge to function

$$f(x) = \int_{x_0}^x g(t)\,dt.$$

Because function g is a uniform limit of continuous functions, it is again a continuous function, thus we have proved all that was needed, see 6.24 about the Riemann integral and a primitive function. □

For infinite series, we can sum up the previous conlusions in this way:

6.43. Corollary. *Consider functions $f_n(x)$ on interval $[a, b]$.*

6.74. 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:

$$\begin{aligned}
\frac{1}{x^3-1} &= \frac{Ax+B}{x^2+x+1} + \frac{C}{x-1},\\
1 &= (Ax+B)(x-1) + C(x^2+x+1),\\
x = 1 &\implies C = \frac{1}{3},\\
x^0 : 1 = C - B &\implies B = -\frac{2}{3},\\
x^2 : 0 = A + C &\implies A = -\frac{1}{3}
\end{aligned}$$

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 $\int \frac{x+2}{x^2+x+1}\,\mathrm{d}x$:

$$\begin{aligned}
&\int \frac{x+2}{x^2+x+1}\,\mathrm{d}x =\\
&= \int \frac{x+\frac{1}{2}}{(x+\frac{1}{2})^2+\frac{3}{4}}\,\mathrm{d}x + \frac{3}{2}\int \frac{1}{(x+\frac{1}{2})^2+\frac{3}{4}}\,\mathrm{d}x =\\
&\quad \left|\begin{array}{l} \text{substitution at the first integral}\\ t = x^2+x+1\\ \mathrm{d}t = 2(x+\frac{1}{2})\,\mathrm{d}x \end{array}\right|\\
&= \frac{1}{2}\int \frac{1}{t}\,\mathrm{d}t + \frac{3}{2}\int \frac{1}{(x+\frac{1}{2})^2+\frac{3}{4}} = \left|\begin{array}{l} \text{substitution at the first integral}\\ s = x+\frac{1}{2}\\ \mathrm{d}s = \mathrm{d}x \end{array}\right|\\
&= \frac{1}{2}\ln(x^2+x+1) + \frac{3}{2}\int \frac{1}{s^2+\frac{3}{4}}\,\mathrm{d}s =\\
&= \frac{1}{2}\ln((x^2+x+1) + \frac{3}{2}\frac{4}{3}\int \frac{1}{\left(\frac{2}{\sqrt{3}}s\right)^2+1}\,\mathrm{d}s =\\
&\quad \left|\begin{array}{l} \text{substitution at the second integral}\\ u = \frac{2}{\sqrt{3}}s\\ \mathrm{d}u = \frac{2}{\sqrt{3}}s\,\mathrm{d}s \end{array}\right|\\
&= \frac{1}{2}\ln(x^2+x+1) + 2\frac{\sqrt{3}}{2}\int \frac{1}{u^2+1}\,\mathrm{d}u =\\
&= \frac{1}{2}\ln(x^2+x+1) + \sqrt{3}\arctan(u) =\\
&= \frac{1}{2}\ln(x^2+x+1) + \sqrt{3}\arctan\left(\frac{2x+1}{\sqrt{3}}\right).
\end{aligned}$$

(1) *If all functions $f_n(x)$ are continuous on $[a, b]$ and the series*

$$S(x) = \sum_{n=1}^\infty f_n(x)$$

uniformly converges to function $S(x)$, then $S(x)$ is continuous on $[a, b]$.

(2) *If all functions $f_n(x)$ are continuously differentiable on interval $[a, b]$, the series $S(x) = \sum_{n=1}^\infty f_n(x)$ converges for some $x_0 \in [a, b]$ and the series $T(x) = \sum_{n=1}^\infty f_n'(x)$ converges uniformly on $[a, b]$, then the series $S(x)$ converges, it is continuously differentiable on $[a, b]$ and $S'(x) = T(x)$, i.e.*

$$\left(\sum_{n=1}^\infty f_n(x)\right)' = \sum_{n=1}^\infty f_n'(x).$$

(3) *If all 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 function $S(x)$ on $[a, b]$, then $S(x)$ is integrable on $[a, b]$ and

$$\int_a^b \left(\sum_{n=1}^\infty f_n(x)\right) dx = \sum_{n=1}^\infty \int_a^b f_n(x)\,dx.$$

6.44. Test of uniform convergence. The simpliest way to find out whether a sequence of functions converges uniformly is a comparison with absolute convergence of a suitable sequence of numbers. This is often called *the Weierstrass test*.

Suppose we have a series of functions $f_n(x)$ on interval $I = [a, b]$ and we have a bound

$$|f_n(x)| \le a_n \in \mathbb{R}$$

for suitable real constants a_n and for all $x \in [a, b]$. We can immediately put a bound on the differences of the partial sums

$$s_k(x) = \sum_{n=1}^k f_n(x)$$

for distinct indices k. For $k > m$ we get

$$|s_k(x) - s_m(x)| = \left|\sum_{n=m+1}^k f_n(x)\right| \le \sum_{n=m+1}^k |f_n(x)| \le \sum_{n=m+1}^k a_k.$$

If a series of (nonnegative) constants $\sum_{n=1}^\infty a_n$ is convergent, then of course the sequence of its partial sums is Cauchy. But we have just verified that in that case the sequence of partial sums $s_n(x)$ will even be uniformly Cauchy.

Thanks to the statement proven above in 6.41 we just proved the following

Theorem (Weierstrass test)**.** *Let $f_n(x)$ be a sequence of functions defined on interval $[a, b]$ with $|f_n(x)| \le a_n \in \mathbb{R}$.*

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.

In total, for the improper integral we can write:

$$
\begin{aligned}
&\int_3^\infty \frac{1}{x^3-1}\,\mathrm{d}x \\
= \ & \frac{1}{3}\lim_{\delta\to\infty}\left[\ln|x-1| - \frac{1}{2}\ln(x^2+x+1) - \sqrt{3}\arctan\left(\frac{2x+1}{\sqrt{3}}\right)\right]_3^\delta \\
= \ & \frac{1}{3}\lim_{\delta\to\infty}\left(\frac{1}{3}\ln|\delta-1| - \frac{1}{2}\ln(\delta^2+\delta+1) - \sqrt{3}\arctan\left(\frac{2\delta+1}{\sqrt{3}}\right)\right) \\
& -\frac{1}{3}\ln 2 + \frac{1}{6}\ln 13 + \frac{\sqrt{3}}{3}\arctan\frac{7}{\sqrt{3}} = \\
= \ & \frac{1}{6}\ln 13 - \frac{1}{3}\ln 2 + \frac{\sqrt{3}}{3}\arctan\frac{7}{\sqrt{3}} - \\
& -\frac{1}{3}\lim_{\delta\to\infty}\ln\left|\frac{x-1}{\sqrt{x^2+x+1}}\right| - \frac{1}{3}\lim_{\delta\to\infty}\sqrt{3}\arctan\left(\frac{2\delta+1}{\sqrt{3}}\right) = \\
= \ & \frac{1}{6}\ln 13 + \frac{1}{\sqrt{3}}\arctan\frac{7}{\sqrt{3}} - \frac{1}{3}\ln 2 - \frac{\sqrt{3}}{6}\pi.
\end{aligned}
$$

□

6.75. 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.

$$
\begin{aligned}
V &= \int_0^2 \frac{x}{2}\,\mathrm{d}x = \pi \\
S &= 2\pi\int_0^2 \sqrt{\frac{x}{2}}\sqrt{1+\frac{1}{8x}}\,\mathrm{d}x = 2\pi\int_0^2\sqrt{\frac{x}{2}+\frac{1}{16}}\,\mathrm{d}x \\
&= \pi\frac{17\sqrt{17}-1}{24}.
\end{aligned}
$$

□

6.76. Compute the area S of a figure composed of two parts of plane bounded by lines $x=0$, $x=1$, $x=4$, the x axis and the graph of a function

$$y = \frac{1}{\sqrt[3]{x-1}}.$$

Solution. First realize that

$$\frac{1}{\sqrt[3]{x-1}} < 0, \quad x\in[0,1), \qquad \frac{1}{\sqrt[3]{x-1}} > 0, \quad x \in (1,4]$$

and

$$\lim_{x\to 1-}\frac{1}{\sqrt[3]{x-1}} = -\infty, \qquad \lim_{x\to 1+}\frac{1}{\sqrt[3]{x-1}} = +\infty.$$

The first part of the figure (below the x axis) is thus bounded by the curves

$$y=0, \quad x=0, \quad x=1, \quad y = \frac{1}{\sqrt[3]{x-1}}$$

with an area given by the improper integral

$$S_1 = -\int\limits_0^1 \frac{1}{\sqrt[3]{x-1}}\,dx;$$

while the second part (above the x axis), which is bounded by the curves

6.45. Consequences for power series. The Weierstrass test is very useful for discussing power series

$$S(x) = \sum_{n=0}^\infty a_n(x-x_0)^n$$

centered at point x_0.

During our first encounter with power series we showed in 5.47 that each such series converges on $(x_0-\delta, x_0+\delta)$, where the so called radius of convergence $\delta \geq 0$ can also be zero or ∞. (see 5.51). In particular, in the proof of the theorem 5.47, we used a comparison with a suitable geometric series to verify the convergence of the series $S(x)$. By the Weierstrass test, the series $S(x)$ converges uniformly on every compact (i.e. finite) interval $[a,b]$ belonging to the interval $(x_0-\delta, x_0+\delta)$. Thus we proved this:

Theorem. *Every power series $S(x)$ is continuous and continuously differentiable at all points inside its interval of convergence. The function $S(x)$ is also integrable and differentiating and integrating can be done term by term.*

In fact, the so called *Abel's theorem* states the power series are continuous even in boundary points of their domain (including eventual infinite limits). We won't prove it here.

Just proven pleasant properties of power series also point at the boundaries of their useablity when simulating dependences of some practical events or processes. In particular, it's not possible to simulate sequentially continuous functions very well by using power series. As we'll see in a moment, for specific needs it's possible to find better sets of functions $f_n(x)$ than the values $f_n(x) = x^n$. The best known examples are the Fourier series and the so called wavelets which we'll discuss in the next chapter.

6.46. Laurent series. In the context of Taylor expansions let's look at a smooth function $f(x) = \mathrm{e}^{-1/x^2}$ from paragraph 6.6. We've seen it's not analytic at zero, because all its derivatives are zero there. So while at all other points x_0 this function is given by convergent Taylor series with radius of convergence $r = |x_0|$, at the origin the series converges at only one point.

But if we substitute the expression $-1/x^2$ for x in the power series for e^x, we get a series of functions

$$S(x) = \sum_{n=0}^\infty \frac{1}{n!}(-1)^n x^{-2n} = \sum_{n=-\infty}^{0}\frac{(-1)^{|n|}}{|n|!}x^{2n},$$

which will converge at all points except for $x \neq 0$ and gives us a good description of behavior near the exceptional point $x = 0$. Thus it seems useful to consider the following more general series quite similar to the power ones:

Laurent series

A series of functions of the form

$$S(x) = \sum_{n=-\infty}^{\infty} a_n(x-x_0)^n$$

is called *a Laurent series centered at* x_0. We call the series convergent if both its parts with positive and negative exponents converge separately.

$$y = 0, \quad x = 1, \quad x = 4, \quad y = \tfrac{1}{\sqrt[3]{x-1}},$$

has an area of

$$S_2 = \int\limits_1^4 \tfrac{1}{\sqrt[3]{x-1}}\, dx.$$

Since

$$\int \tfrac{1}{\sqrt[3]{x-1}}\, dx = \tfrac{3}{2}\sqrt[3]{(x-1)^2} + C,$$

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 $\delta > 0$. □

6.77. Determine the avarage velocity v_p of a solid in the time interval $[1, 2]$, if its velocity is

$$v(t) = \tfrac{t}{\sqrt{1+t^2}}, \qquad t \in [1, 2].$$

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

$$v_p = \tfrac{1}{2-1}\int\limits_1^2 \tfrac{t}{\sqrt{1+t^2}}\, dt = \int\limits_2^5 \tfrac{1}{2\sqrt{x}}\, dx = \sqrt{5} - \sqrt{2},$$

with $1 + t^2 = x$, $t\, dt = dx/2$. □

6.78. Compute the length s of a part of the curve calles tractrix given by the parametric description

$$f(t) = r\cos t + r\ln\left(\operatorname{tg}\tfrac{t}{2}\right), \qquad g(t) = r\sin t, \qquad t \in [\pi/2, a],$$

where $r > 0$, $a \in (\pi/2, \pi)$.

Solution. Since

$$f'(t) = -r\sin t + \tfrac{r}{2\operatorname{tg}\frac{t}{2}\cdot\cos^2\frac{t}{2}} = -r\sin t + \tfrac{r}{\sin t} = \tfrac{r\cos^2 t}{\sin t},$$

$g'(t) = r\cos t$ on interval $[\pi/2, a]$,

for the length s we get

$$s = \int\limits_{\pi/2}^{a} \sqrt{\tfrac{r^2\cos^4 t}{\sin^2 t} + r^2\cos^2 t}\; dt = \int\limits_{\pi/2}^{a} \sqrt{\tfrac{r^2\cos^2 t}{\sin^2 t}}\; dt =$$

$$-r\int\limits_{\pi/2}^{a} \tfrac{\cos t}{\sin t}\, dt = -r\left[\ln(\sin t)\right]_{\pi/2}^{a} = -r\ln(\sin a).$$

□

6.79. 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

The purpose of Laurent series can be seen at rational functions. Consider such function $S(x) = f(x)/g(x)$ with coprime polynomials f and g and consider a root x_0 of polynomial $g(x)$. If the multiplicity of this root is s, then after multiplication we get function $\tilde{S}(x) = S(x)(x-x_0)^s$, which will now be analytic on some neighbouthood of the point x_0 and therefore we can write

$$\begin{aligned} S(x) &= \frac{a_{-s}}{(x-x_0)^s} + \dots + \frac{a_{-1}}{x-x_0} + a_0 + a_1(x-x_0) + \dots \\ &= \sum_{n=-s}^{\infty} a_n(x-x_0)^n. \end{aligned}$$

Now consider seperate parts

$$S(x) = S_- + S_+ = \sum_{n=-\infty}^{-1} a_n(x-x_0)^n + \sum_{n=0}^{\infty} a_n(x-x_0)^n.$$

As for the series S_+, Theorem 5.47 implies that its radius of convergence R is given by the equality

$$R^{-1} = \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

If we apply the same idea to the series S_- with $1/x$ subtituted for x though, we'll find out the series $S_-(x)$ converges for $|x - x_0| > r$, where

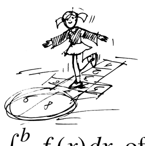
$$r^{-1} = \limsup_{n\to\infty} \sqrt[n]{|a_{-n}|}.$$

These notions remain completely true even for complex values of x substituted into our expressions.

Theorem. *A 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 $|x - x_0| > R$.*

Hence we can see the Laurent series need not converge at any point at all, because we can have values $R < r$. But if we look for example at 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 will really converge in the punctured neighbouthood of this point x_0, while R will be given exactly by the distance to the closest root of the denominator. In case of our first example, for the function e^{-1/x^2} we have $r = 0$ and $R = \infty$.

6.47. Numerical approximation of integration. Just like at the end of the previous part of the text (see paragraph 6.17), we'll now use the Taylor expansion to propose as good and simple approximations of integration as possible. We'll work with an integral $I = \int_a^b f(x)dx$ of analytic function $f(x)$ and a uniform partition of the interval $[a, b]$ using points $a = x_0, x_1, \dots, x_n = b$ with distances $x_i - x_{i-1} = h > 0$. We'll denote the points in the middle of the intervals in the partitions by $x_{i+1/2}$ and the values of our function at the points of the partition by $f(x_i) = f_i$.



We'll compute the contribution of one segment of the partition to the integral by the Taylor expansion and the previous theorem. We intentionally integrate symmetrically around the middle values so that the derivatives of odd orders cancel each other out while 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

$$\int\limits_0^3 \pi f^2(x)\,dx = \pi \int\limits_0^3 \sqrt{9x^2 - x^4}\,dx.$$

Using the substitution $t = \sqrt{9 - x^2}$ ($xdx = -tdt$), we can easily compute

$$\int\limits_0^3 \sqrt{9x^2 - x^4}\,dx = \int\limits_0^3 x \cdot \sqrt{9 - x^2}\,dx = -\int\limits_3^0 t^2\,dt = 9,$$

and receive the result 18π. □

6.80. 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

$$V = \pi \int\limits_a^{+\infty} \left(\tfrac{1}{x}\right)^2 dx = \pi \int\limits_a^{+\infty} \tfrac{1}{x^2}\,dx = \pi \left(\lim_{x\to+\infty} -\tfrac{1}{x} - \left(-\tfrac{1}{a}\right)\right) = \tfrac{\pi}{a}$$

and

$$S = 2\pi \int\limits_a^{+\infty} \tfrac{1}{x} \cdot \sqrt{1 + \left(-\tfrac{1}{x^2}\right)^2}\,dx = 2\pi \int\limits_a^{+\infty} \tfrac{\sqrt{x^4+1}}{x^3}\,dx \geq 2\pi \int\limits_a^{+\infty} \tfrac{1}{x}\,dx =$$

$$2\pi \left(\lim_{x\to+\infty} \ln x - \ln a \right) = +\infty.$$

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 Torriccelli's paradox. But realize that a real color painting has a 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. □

Another problems about computing lengths of curves, areas of plane figures and volumes of parts of space can be found on page 413.

6.81. Apllications of the integral criterion of convergence. Now let's get back to (number) series. Thanks to the integral criterion of convergence (see 6.33), we can decide the question of convergence for a wider class of series: Decide, whether the following sums converge of diverge:

a) $\sum\limits_{n=1}^{\infty} \frac{1}{n \ln n}$,

b) $\sum\limits_{n=1}^{\infty} \frac{1}{n^2}$.

Solution. First notice, that we cannot decide the convergence of none of these series by using the ratio or root test (all limits $\lim\limits_{n\to\infty} |\frac{a_{n+1}}{a_n}|$ and $\lim\limits_{n\to\infty} \sqrt[n]{a_n}$ equal 1). Using the integral criterion for convergence of series, we obtain:

a)

$$\int_1^\infty \frac{1}{x\ln(x)}\,\mathrm{d}x = \int_0^\infty \frac{1}{t}\,\mathrm{d}t = \lim_{\delta\to\infty} [\ln(t)]_0^\delta = \infty,$$

integrating:

$$\begin{aligned}\int_{-h/2}^{h/2} f(x_{i+1/2} + t)dt &= \int_{-h/2}^{h/2} \left(\sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_{i+1/2}) t^n \right) dt \\ &= \sum_{k=0}^{\infty} \left(\int_{-h/2}^{h/2} \frac{1}{k!} f^{(k)}(x_{i+1/2}) t^k\,dt \right) \\ &= \sum_{k=0}^{\infty} \frac{h^{2k+1}}{2^{2k}(2k+1)!} f^{(2k)}(x_{i+1/2}).\end{aligned}$$

A very simple numerical approximation of integration on one segment of the partition is the so called *trapezoidal rule*, which 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

$$P_i = \frac{1}{2}(f_i + f_{i+1})h$$

so in total we can approximate the integral I by value

$$I_{\text{lich}} = \sum_{i=0}^{n-1} P_i = \frac{h}{2}(f_0 + 2f_1 + \cdots + 2f_{n-1} + f_n).$$

We'll now compare I_{lich} with the exact value of I computed by contributions over seperate segments of the partition. We can express the values f_i by middle values and derivatives $f^{(k)}_{i+1/2}$ in this way:

$$\begin{aligned}f_{i+1/2\pm1/2} = f_{i+1/2} \pm \frac{h}{2} f'_{i+1/2} &+ \frac{h^2}{2!2^2} f''(i+1/2) \\ &\pm \frac{h^3}{3!2^3} f^{(3)}(i+1/2) + \ldots,\end{aligned}$$

so for the contribution P_i to the approximation we get

$$P_i = \frac{1}{2}(f_i + f_{i+1})h = h\big(f_{i+1/2} + \frac{h^2}{2!2^2} f''(i+1/2)\big) + O(h^5).$$

>From here we get an estimations of the error $I - I_{\text{lich}}$ over one segment of the partition

$$\begin{aligned}\Delta_i &= h\big(f_{i+1/2} + \frac{h^2}{24} f''_{i+1/2} - f_{i+1/2} - \frac{h^2}{8} f''_{i+1/2} + O(h^4)\big) \\ &= \frac{h^3}{12} f''_{i+1/2} + O(h^5).\end{aligned}$$

The total error is thus estimated as

$$I - I_{\text{lich}} = \frac{1}{12} n h^3 f'' + n\,O(h^5) = \frac{1}{12}(b-a)h^2 f'' + O(h^4)$$

where f'' represents the approximation of the second derivative of f.

If the linear approximation of the function over the seperate segments doesn't suffice, we can try can an approximation by a quadratic polynomial. To do that, we'll always need three points, so we'll work with segments of the partition in pairs. Suppose $n = 2m$ and consider x_i with odd indices. We'll require

$$f_{i+1} = f(x_i + h) = f_i + \alpha h + \beta h^2$$

$$f_{i-1} = f(x_i - h) = f_i - \alpha h + \beta h^2$$

which gives (see the similarity to the difference for approximating the second derivative)

$$\beta = \frac{1}{2h^2}(f_{i+1} + f_{i-1} - 2f_i).$$

hence the given series diverges.

b)

$$\int_1^\infty \frac{1}{x^2}\,\mathrm{d}x = \lim_{\delta\to\infty}\left[-\frac{1}{x}\right]_1^\delta = 1,$$

hence the given series converges.

□

6.82. Using the integral criterion, decide the convergence of series

$$\sum_{n=1}^{\infty} \tfrac{1}{(n+1)\ln^2(n+1)}.$$

Solution. The function

$$f(x) = \tfrac{1}{(x+1)\ln^2(x+1)}, \qquad x \in [1,+\infty)$$

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

$$\int\limits_1^{+\infty} \tfrac{1}{(x+1)\ln^2(x+1)}\,dx = \int\limits_{\ln 2}^{+\infty} \tfrac{1}{y^2}\,dy = \tfrac{1}{\ln 2}.$$

Hence the series converges. □

G. Uniform convergence

6.83. Does the sequence of functions

$$y_n = \mathrm{e}^{\frac{x^4}{4n^2}}, \qquad x \in \mathbb{R}, \qquad n \in \mathbb{N}$$

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

$$\lim_{n\to\infty} \mathrm{e}^{\frac{x^4}{4n^2}} = \mathrm{e}^0 = 1, \qquad x \in \mathbb{R}.$$

But the computation

$$y_n\left(\sqrt{2n}\right) = \mathrm{e} > 2 \quad \text{for all } n \in \mathbb{N}$$

implies that it's not a uniform convergence. (In the definition of uniform convergence, it suffices to consider $\varepsilon \in (0,1)$.) □

6.84. Decide whether the series

$$\sum_{n=1}^{\infty} \tfrac{\sqrt{x}\cdot n}{n^4+x^2}$$

converges uniformly on the interval $(0,+\infty)$.

Solution. Using the denotation

$$f_n(x) = \tfrac{\sqrt{x}\cdot n}{n^4+x^2}, \qquad x > 0, \qquad n \in \mathbb{N},$$

we have

$$f_n'(x) = \tfrac{n\left(n^4-3x^2\right)}{2\sqrt{x}\left(n^4+x^2\right)^2}, \qquad x > 0, \qquad n \in \mathbb{N}.$$

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

$$f_n\left(\tfrac{n^2}{\sqrt{3}}\right) = \tfrac{\sqrt[4]{27}}{4n^2} \quad \text{a} \quad \sum_{n=1}^{\infty} \tfrac{\sqrt[4]{27}}{4n^2} = \tfrac{\sqrt[4]{27}}{4}\sum_{n=1}^{\infty}\tfrac{1}{n^2} < +\infty,$$

The area of approximation of the integral over two segments of the partition between x_{i-1} and x_{i+1} is now estimated by the expression

$$\begin{aligned} P_i &= \int_{-h}^{h} f_i + \alpha t + \beta t^2\,dt = 2hf_i + \frac{2}{3}\beta h^3 \\ &= 2hf_i + \frac{2h}{6}(f_{i+1} + f_{i-1} - 2f_i) \\ &= \frac{h}{3}(4f_{i+1} + f_{i-1} - 2f_i). \end{aligned}$$

This procedure is called *Simpson's rule*. The whole integral is now approximated by

$$I_{\text{Simp}} = \frac{1}{3}h\big(f_0 + f_{2n} + 4\sum_{\text{odd } k} f_k + 2\sum_{\text{even } k} f_k\big).$$

Similarly to the procedure above we can derive that 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 approximation of the fourth derivative of f.

By the end of this chapter, we'll stop at other concepts of integration. First we'll show a modification of the Riemann integral, which will later be useful in notions about probability and statistics. We'll mostly stay in an area of notions and comments though, readers interested in a thorough explication will need to find another sources.

6.48. Riemann–Stieltjes integral. In our idea of integration as summing 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 would take the increments with different weights. This could be surely arranged at the infinitesimal level by interchanging the differential dx for $\varphi(x)dx$ for some suitable function φ. We've seen this behavior for example while computing the length of a parametrized curve in space.

Surely we can also imagine that at some point x_0, the increment of the integrated quantity is given by $\alpha f(x_0)$ independently on the size of the increment of x. For example we can observe the probability that the amount of per mille of alcohol in blood of a driver at a test will be at most x. With quite a large probability we'll obtain value 0, thus for any integral sum, the segment containing zero must contribute by a constant nonzero contribution, independent on the norm of the partition. We cannot simulte such behavior by multiplying the differential dx by some real function. Instead we can generalize the Riemann integral in this way:

Choose a real nondecreasing function g on a finite interval $[a,b]$. For every partition Ξ with representants ξ_i and points of partition

$$a = x_0, x_1, \dots, x_n = b$$

we define the *Riemann–Stieltjes integral sum* of function $f(x)$ as

$$S_\Xi = \sum_{i=1}^{n} f(\xi_i)\big(g(x_i) - g(x_{i-1})\big).$$

Then we say the Riemann–Stieltjes integral

$$I = \int_a^b f(x)dg(x)$$

according to the Weierstrass test, the series $\sum_{n=1}^{\infty} f_n(x)$ converges uniformly on the interval $(0, +\infty)$. □

6.85. For $x \in [-1, 1]$, add

$$\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n(n+1)}\, x^{n+1}.$$

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

$$\textstyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)}\, x^{n+1} = \sum_{n=1}^{\infty} \left(\frac{(-1)^{n+1}}{n} \int x^n\, dx \right) =$$

$$\boxed{\int \sum_{n=1}^{\infty} \left(\frac{(-1)^{n+1}}{n}\, x^n \right) dx} = \textstyle\int \sum_{n=1}^{\infty} \big((-1)^{n+1} \int x^{n-1}\, dx \big)\, dx =$$

$$\textstyle\int \big(\int \sum_{n=1}^{\infty} (-x)^{n-1}\, dx \big)\, dx = \int \big(\int 1 - x + x^2 - x^3 + \cdots\, dx \big)\, dx =$$

$$\textstyle\int \big(\int \frac{1}{1+x}\, dx \big)\, dx = \boxed{\displaystyle\int \ln(1+x) + C_1\, dx}.$$

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

$$\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n}\, x^n = \ln(1+x) + C_1, \qquad x \in (-1, 1).$$

The choice $x = 0$ then yields $0 = \ln 1 + C_1$, i.e. $C_1 = 0$. Next,

$$\textstyle\int \ln(1+x)\ dx = \big|\text{ per partes }\big| = \begin{vmatrix} u = \ln(1+x) & u' = \frac{1}{1+x} \\ v' = 1 & v = x \end{vmatrix} =$$

$$x \ln(1+x) - \textstyle\int \frac{x}{1+x}\, dx = x \ln(1+x) - \int 1 - \frac{1}{1+x}\, dx =$$
$$x \ln(1+x) - x + \ln(1+x) + C_2 = (x+1) \ln(x+1) - x + C_2.$$

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

$$\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n(n+1)}\, x^{n+1} = (x+1)\ \ln(x+1) - x, \qquad x \in (-1, 1).$$

Moreover, according to Abel's theorem (see 6.45), 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:

$$\lim_{x \to -1^+} (x+1)\ \ln(x+1) - x = \lim_{t \to 0^+} t\ \ln t + 1$$
$$= \lim_{t \to 0^+} \frac{\ln t}{\frac{1}{t}} + 1 = \lim_{t \to 0^+} \frac{\frac{1}{t}}{-\frac{1}{t^2}} + 1 = \lim_{t \to 0^+} -t + 1 = 1.$$

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 lesser than δ, we have

$$|S_\Xi - I| < \varepsilon.$$

For example, if we choose $g(x)$ on interval $[0, 1]$ as a sequentially constant function with finitely many discontinuities $c_1, \dots, c_k$ 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

$$I = \int_0^1 f(x) dg(x) = \sum_{i=1}^{k} \alpha_i f(c_k).$$

By the same technique we used for the Riemann integral, we can now define upper and lower sums and uppoer and lower Riemann–Stieltjes integral, which have the advantage that for bounded functions they always exist and their values coincide if and only if the Riemann–Stieltjes integral in the above sense exists.

We already encountered problems with Riemann integration of functions that were "too jumpy". Technically, for function $g(x)$ on a finite interval $[a, b]$ we define its *variation* by

$$\operatorname{var}_a^b g = \sup_{\Xi} \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})|,$$

where we take the supremum over all partitions Ξ of the interval $[a, b]$. If the supremum is infinite, we say $g(x)$ has an unbounded variation on $[a, b]$, otherwise we say g is a function with a bounded variation on $[a, b]$.

Similarly to the procedure for the Riemann integral, we can quite easily derive the following:

Theorem. *Let $f(x)$ and $g(x)$ be real functions on a finite interval $[a, b]$.*

(1) If $g(x)$ is decreasing and continuously differentiable, then the Riemann integral on the left side and the Riemann–Stieltjes integral on the right side both exist simultaneously and 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 nondecreasing function with a finite variation, then the integral $\int_a^b f(x) dg(x)$ exists.

6.49. Kurzweil integral. The last stop will be a modification of the Riemann integral, which fixes the unfortunate behavior at the third point in the paragraph 6.37, i.e. the limits of the nondecreasing sequences of integrable functions will again be integrable. Then we will be able to interchange the order of the limit process and integration in these cases, just like with uniform convergence.

First notice what's the essence of the problem. Intuitively we should assume that very small sets must have a zero size, and thus the changes of values of the functions on such sets shouldn't influence the integration. Moreover, a countable union of such "negligible for the purpose of integration" sets should have a zero size again. Surely we would expect that for example the set of rational

Of course, the convergence of the series at points ±1 can be verified directly. It's even possible to directly deduce that $\sum\limits_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$ (by writing out $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$. □

6.86. Sum of a series. Using theorem 6.41 "about the interchange of a limit and an integral of a sequence of uniformly convergent functions", we'll now add the number series

$$\sum_{n=1}^{\infty} \frac{1}{n2^n}.$$

We'll use the fact that $\int\limits_{2}^{\infty} \frac{\mathrm{d}x}{x^{n+1}} = \frac{1}{n2^n}$.

Solution. On interval $(2, \infty)$, the series of functions $\sum_{n=1}^{\infty} \frac{1}{x^{n+1}}$ converges uniformly. That is implied for example by the Weierstrass test: each of the function $\frac{1}{x^{n+1}}$ is decreasing on interval $(2, \infty)$, thus their values are at most $\frac{1}{2^{n+1}}$; the series $\sum_{n=1}^{\infty} \frac{1}{2^{n+1}}$ is convergent though (it's a geometric series with quotient $\frac{1}{2}$). Hence according to the Weierstrass test, the series of functions $\sum_{n=1}^{\infty} \frac{1}{x^{n+1}}$ converges uniformly. We can even write the resulting function explicitly. Its value at any $x \in (2, \infty)$ is the value of the geometric series with quotient $\frac{1}{x}$, so if we denote the limit by $f(x)$, we have

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{x^{n+1}} = \frac{1}{x^2}\frac{1}{1-\frac{1}{x}} = \frac{1}{x(x-1)}.$$

By using (6.43) (3), we get

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{1}{n2^n} &= \sum_{n=1}^{\infty} \int_2^{\infty} \frac{\mathrm{d}x}{x^{n+1}} \\
&= \int_2^{\infty} \left(\sum_{n=1}^{\infty} \frac{1}{x^{n+1}} \right) \mathrm{d}x \\
&= \int_2^{\infty} \frac{1}{x(x-1)} \,\mathrm{d}x \\
&= \lim_{\delta\to\infty} \int_2^{\delta} \frac{1}{x-1} - \frac{1}{x} \,\mathrm{d}x \\
&= \lim_{\delta\to\infty} [(\ln(\delta - 1) - \ln(\delta) - \ln(1) + \ln 2] \\
&= \lim_{\delta\to\infty} \left[\ln\left(\frac{\delta-1}{\delta} \right) \right] + \ln(2) = \ln\left(\lim_{\delta\to\infty} \frac{\delta-1}{\delta} \right) + \ln 2 \\
&= \ln 2
\end{aligned}$$

□

6.87. Consider function $f(x) = \sum_{n=1}^{\infty} ne^{-nx}$. Determine

$$\int_{\ln 2}^{\ln 3} f(x) \,\mathrm{d}x.$$

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,

numbers inside a finite interval would have this property, hence its characteristic function should be integrable and the value of such interval should be zero.

We say the set $A \subset \mathbb{R}$ has a *zero measure*, if for every $\varepsilon > 0$ we can find a covering of set A by a countable system of open intervals J_i, $i = 1, 2, \dots$ such that

$$\sum_{i=1}^{\infty} m(J_i) < \varepsilon.$$

In the following, by the statement "function f has the given property on set B almost everywhere" we'll always mean the fact 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 sequentially continuous function is continuous almost everywhere etc.

We'd now like to modify the definition of the Riemann integral so that when choosing the partition and the corresponding Riemann sums, we would be able to eliminate the omnious effect of the values of the integrated function on a before known set of zero measure. It also seems reasonable to try to guarantee that the segments in the given partitions with representants would have the property that near points of such set, they would be controllably small.

A positive real function δ on a finite interval $[a, b]$ is called a *calibre*. We call a partition Ξ of interval $[a, b]$ with representants ξ_i δ–calibrated, if we have

$$\xi_i - \delta(\xi_i) < x_{i-1} \le \xi_i \le x_i < \xi_i + \delta(\xi_i)$$

for all i.

For further procedure, it's essential to verify that for every calibre δ, a δ–calibrated partition with representants can be found. This statement is called Cousin's lemma and can be proven for example in the usual way based upon the properties of supremas. For a given calibre δ on $[a, b]$, we'll denote by M the set of all points $x \in [a, b]$ such that a δ–calibrated partition with representants can be found on $[a, x]$. Surely M is nonempty and bounded, thus it has a supremum s. If $s \ne b$, then we could find a calibrated partition with a representant at s, which leads to a contradiction.

Now we can define a generalization of the Riemann integral in this way:

Definition. Function f defined on a finite interval $[a, b]$ has its Kurzweil integral

$$I = \int_a^b f(x)\,dx,$$

if for every $\varepsilon > 0$, there exists a calibre δ such that for every δ–calibrated partition with representants Ξ, the inequality $|S_\Xi - I| < \varepsilon$ holds for the corresponding Riemann sum S_Ξ.

6.50. Properties of the Kurzweil integral. First notice that when defining the Kurzweil integral, we only bounded the set of all partitions, for which we take the Riemann sums into account. Hence if our function is Riemann integrable, then it must also have the Kurzweil integral and these two integrals are equal.

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 (6.43) (3), we have

$$\begin{aligned}\int_{\ln 2}^{\ln 3} f(x)\,\mathrm{d}x &= \int_{\ln 2}^{\ln 3}\sum_{n=1}^{\infty} ne^{-nx} = \\ &= \sum_{n=1}^{\infty}\int_{\ln 2}^{\ln 3} ne^{-nx}\,\mathrm{d}x = \\ &= \sum_{n=1}^{\infty}[-e^{-nx}]_{\ln 2}^{\ln 3} = \sum_{n=1}^{\infty}\left(\frac{1}{2^n}-\frac{1}{3^n}\right) = 1-\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

□

6.88. Determine the following limit (give reasons for the procedure of computation):

$$\lim_{n\to\infty}\int_0^{\infty}\frac{\cos\left(\frac{x}{n}\right)}{\left(1+\frac{x}{n}\right)^n}\,\mathrm{d}x.$$

Solution. First we'll determine $\lim\limits_{n\to\infty}\frac{\cos(\frac{x}{n})}{(1+\frac{x}{n})^n}$. The sequence of these functions converges pointwise and we have

$$\lim_{n\to\infty}\frac{\cos(\frac{x}{n})}{\left(1+\frac{x}{n}\right)^n} = \frac{1}{\lim\limits_{n\to\infty}\left(1+\frac{x}{n}\right)^n} \overset{(\|\mathbf{??}\|)}{=} \frac{1}{e^x}$$

It can be shown that the given sequence converges uniformly. Then according to (6.41) ,

$$\begin{aligned}\lim_{n\to\infty}\int_0^{\infty}\frac{\cos\left(\frac{x}{n}\right)}{\left(1+\frac{x}{n}\right)^n}\,\mathrm{d}x &= \int_0^{\infty}\left[\lim_{n\to\infty}\frac{\cos\left(\frac{x}{n}\right)}{\left(1+\frac{x}{n}\right)^n}\right]\mathrm{d}x = \\ &= \int_0^{\infty}\frac{1}{e^x} = 1\end{aligned}$$

We leave the verification of uniform convergence to the reader (we only point out that the discussion is more complicated than in the previous cases).

□

For the same reason, we can repeat the argumentation in Theorem 6.24 about simple properties of the Riemann integral and again verify that the Kurzweil 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. For proving this, it only suffices to think through little modifications when discussing the refined partitions, which moreover should be δ–calibrated.

Analogously, for the case of monotonic sequences of pointwise convergent functions, we can extend the argumentation verifying that the limits of uniformly convergent sequences of integrable functions f_n are again integrable and the integral of the limit is the limit of the values of integrals f_n.

Finally, the Kurweil integral behaves in the way we would like it to, even to sets with zero measure:

Theorem. *Consider a function f on interval $[a,b]$, which is zero almost everywhere. Then the Kurzweil integral $\int_a^b f(x)d(x)$ exists and equals zero.*

Proof. This is a nice illustration of the idea that we can get rid of the influence of values on a small set by a smart choice of calibre. Denote by M the corresponding set of zero measure, outside of which $f(x)=0$ and write $M_k \subset [a,b]$, $k=1,\dots,$ for the subset of the points for which $k-1 < |f(x)| \le k$. Because all the sets M_k have zero measure, we can cover it by a countable system of in sum arbitrarily small and pairwise disjoint open intervals $J_{k,i}$. Now definte the calibre $\delta(x)$ for $x\in J_{k,i}$ so that the whole intervals $(x-\delta(x), x+\delta(x))$ were still contained in $J_{k,i}$. Outside of set M, we then define δ arbitrarily.

For δ–calibrated partition Ξ of the interval $[a,b]$ we can then put a bound on the corresponding Riemann sum

$$\begin{aligned}\left|\sum_{j=0}^{n-1} f(\xi_n)(x_{i+1}-x_i)\right| &= \left|\sum_{\substack{j=0\\ \xi_i\in M}}^{n-1} f(\xi_n)(x_{i+1}-x_i)\right| \\ &\le \sum_{k=1}^{\infty}\sum_{\substack{j=0\\ \xi_i\in M_k}}^{n-1}\left|f(\xi_n)\right|(x_{i+1}-x_i) \\ &\le \sum_{k=1}^{\infty} k\Bigg(\sum_{\substack{j=0\\ \xi_i\in M_k}}^{n-1} m(J_{k,j})\Bigg)\end{aligned}$$

Hence if we want this bound to be smaller than ε for an ε known in advancem it suffices to choose the covering by the intervals $J_{k,j}$ so that

$$\sum_{j=1}^{\infty} m(J_{k,j}) = \frac{\varepsilon}{k2^k}.$$

Then in the last expression we can substitute for the inner sum, add the geometric series $\sum_{k=1}^{\infty} 2^{-k}$ and get exactly the required ε. □

Corollary. *We don't change the Kurzweil integrability of a given function $f(x)$ neither the value of its integral if we change the values $f(x)$ on a set of zero measure.*

6.51. Comments about the integration. To finish ....

absolutely continuous functions, the relation between the indefinite integral and the primitive function, intergration in absolute value, Lebesgue integral

H. Expletory examples for the whole chapter

6.89. 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? ○

6.90. Determine the total course of function

$$f(x) = -\tfrac{x^2}{x+1}, \quad x \in \mathbb{R} \setminus \{-1\}.$$

Hence determine (if sensible):

(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.

○

6.91. Determine the course of the function

$$f(x) = \tfrac{1-x^3}{x^2}.$$

By determining the course of function f (not only in this example) we mean „determining the domain, the range and eventual parity or periodicity; computing the limits

$$\lim_{x\to-\infty} f(x) \quad \text{a} \quad \lim_{x\to+\infty} f(x),$$

if they exist; determining the discontinuities and their kind including the according one-sided limits (if they exist), zeros (if they exist) and the intervals where the function is positive and where it's negative; determining the first (and the second, if needed) derivative and *the intervals on which the function increases, decreases or remains constant*; finding the stationary (critical) points and all *local extremes* (if they exist); determining *the points of inflection* and *the intervals on which the function is convex and concave*; computing the values at significant points (i.e. find the values of the function at stationary points and points of inflection, if it helps when plotting the graph, and find the points of intersection with the axes, if they exist); plot its *graph* with *the asymptotes*“. ○

6.92. Determine the course of the function

$$f(x) = \tfrac{x^3-3x^2+3x+1}{x-1}.$$

○

6.93. Determine the course of the function

$$f(x) = \sqrt[3]{x}\,\mathrm{e}^{-x}.$$

○

6.94. Determine the course of the function

$$f(x) = \operatorname{arctg} \tfrac{x}{2-x}.$$

○

6.95. Determine the course of the function $\frac{\ln x}{x}$; among other things, find the extremes, the points of inflection and the asymptotes and plot its graph. ○

6.96. Determine the course of the function; among other things, find the extremes, the points of inflection and the asymptotes:

$$\ln(x^2 - 3x + 2) + x.$$

○

6.97. Determine the course of the function; among other things, find the extremes, the points of inflection and the asymptotes:

$$(x^2 - 2)e^{x^2-1}.$$

○

6.98. Determine the course of the function; among other things, find the extremes, the points of inflection and the asymptotes:

$$\ln(2x^2 - x - 1).$$

○

6.99. Determine the course of the function (**among other things,** find the extremes, the points of inflection and the asymptotes):

$$\frac{x^2 - 2}{x - 1}.$$

○

6.100. Using the basic formulas, determine any primitive function to 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) $y = \frac{\cos x}{1+\sin x}, \quad x \in \left(-\frac{\pi}{2}, \frac{3\pi}{2}\right).$

○

6.101. 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) $y = \frac{1}{\sin^2 x\, \cos^2 x}, \quad x \in \left(0, \frac{\pi}{2}\right).$

○

6.102. Find the primitive function to function

$$y = \mathrm{e}^x + \tfrac{3}{\sqrt{4-x^2}}$$

na intervalu $(-2, 2)$. ○

6.103. Determine

$$\int \tfrac{x^3}{1+x^4}\, dx, \quad x \in \mathbb{R}.$$

○

6.104. Determine

$$\int \frac{4}{x^2-2x+3}\, dx, \quad x \in \mathbb{R}.$$

○

6.105. For $x \in (0, 1)$, compute

$$\int \left(\frac{x^2+1}{x\left(x^2-1\right)} + \frac{3}{\sqrt{4-4x^2}} + 4\sin x - 5\cos x \right) dx\,.$$

○

6.106. 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. ○

6.107. By repeated use of integration by parts, for all $x \in \mathbb{R}$ determine

(a) $\int x^2 \sin x \; dx;$
(b) $\int x^2 \,\mathrm{e}^x\, dx.$

○

6.108. For example by using integration by parts, determine

$$\int x \ln^2 x \; dx$$

for $x > 0$. ○

6.109. Using integration by parts, determine

$$\int \left(2 - x^2\right) \mathrm{e}^x\, dx$$

on the whole real line. ○

6.110. 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) $\int \mathrm{e}^{-x^3} x^2\, dx, \quad x \in \mathbb{R};$
(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) $\int \frac{\mathrm{e}^x}{\mathrm{e}^{2x}+3}\, dx, \quad x \in \mathbb{R};$
(h) $\int \sin\sqrt{x}\; dx, \quad x > 0$

by using the substitution method. ○

6.111. 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. ○

6.112. For $x \in (-\pi/2, \pi/2)$ compute

$$\int \frac{dx}{1+\sin^2 x}$$

using the substitution $t = \operatorname{tg} x$. ○

6.113. Determine

$$\int \frac{\sqrt{x}}{\sqrt{x}+1}\, dx\,, \quad x > 0.$$

in an arbitrary way. ○

6.114. Compute

(a) $\int x^n \ln x \; dx, \quad x > 0, \; n \neq -1;$
(b) $\int \frac{x}{1+x^4} \, dx, \quad x \in \mathbb{R}.$

○

6.115. 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 subtitution $x = t^s$, 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 $a + bx^n \; = t^s \, x^n$, 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)

$$\int \frac{(2+5x)^3}{\sqrt[4]{x^3}} \, dx = \int x^{-\frac{3}{4}} \, (2 + 5x)^3 \, dx = \left| \begin{array}{c} p \in \mathbb{Z} \\ x = t^4 \\ dx = 4t^3 \; dt \end{array} \right| = 4 \int \left(2 + 5t^4\right)^3 \; dt =$$
$$4 \int \left(8 + 60t^4 \; + 150t^8 \; + 125t^{12}\right) dt = 4 \left(8t \; + 12t^5 \; + \tfrac{50}{3} \, t^9 \; + \tfrac{125}{13} \, t^{13}\right) + C =$$
$$4 \left(8\sqrt[4]{x} + 12\sqrt[4]{x^5} + \tfrac{50}{3} \, \sqrt[4]{x^9} + \tfrac{125}{13} \, \sqrt[4]{x^{13}}\right) + C;$$

(b)

$$\int \frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}} dx = \int x^{-\frac{1}{2}} \, \left(1 + x^{\frac{1}{4}}\right)^{\frac{1}{3}} \, dx = \left| \begin{array}{c} p \notin \mathbb{Z}, \; \frac{m+1}{n} \in \mathbb{Z} \\ 1 + x^{\frac{1}{4}} \; = t^3 \\ x = (t^3 \; - 1)^4 \\ dx = 12t^2 \, \left(t^3 \; - 1\right)^3 \; dt \end{array} \right| = 12 \int t^3 \, \left(t^3 \; - 1\right) \, dt =$$
$$12 \int t^6 \; - t^3 \; dt = 12 \left(\tfrac{t^7}{7} - \tfrac{t^4}{4}\right) + C = 12\sqrt[3]{(1 + \sqrt[4]{x})^4} \left(\tfrac{1+\sqrt[4]{x}}{7} - \tfrac{1}{4}\right) + C;$$

(c)

$$\int \frac{1}{\sqrt[4]{1+x^4}} \, dx = \int \left(1 + x^4\right)^{-\frac{1}{4}} \, dx = \left| \begin{array}{c} p \notin \mathbb{Z}, \; \frac{m+1}{n} \notin \mathbb{Z}, \; \frac{m+1}{n} + p \in \mathbb{Z} \\ 1 + x^4 \; = t^4 x^4 \\ x = \left(t^4 \; - 1\right)^{-\frac{1}{4}} \\ dx = -t^3 \, \left(t^4 \; - 1\right)^{-\frac{5}{4}} \; dt \end{array} \right| = - \int \frac{t^2}{t^4 - 1} dt =$$
$$- \int \frac{t^2}{(t-1)(t+1)(t^2+1)} \, dt = -\tfrac{1}{4} \int \left(\tfrac{1}{t-1} - \tfrac{1}{t+1} + \tfrac{2}{t^2+1}\right) dt =$$
$$-\tfrac{1}{4} \left(\ln |\, t - 1 \,| - \ln |\, t + 1 \,| + 2 \operatorname{arctg} t\right) \, + C =$$
$$-\frac{1}{4} \left[\ln \frac{\sqrt[4]{\frac{1}{x^4}+1}-1}{\sqrt[4]{\frac{1}{x^4}+1}+1} + 2 \operatorname{arctg} \left(\sqrt[4]{\frac{1}{x^4} + 1} \right) \right] + C.$$

□

6.116. 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 $\int f(\sin x, \cos x)\ dx$ 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 hold, the substitution $t = \operatorname{tg}\frac{x}{2}$ 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

$$\int \frac{\sin^3 x}{1+4\cos^2 x+3\sin^2 x}\,dx = \int \frac{\sin x(1-\cos^2 x)}{4+\cos^2 x}\,dx = \int \frac{-(1-t^2)}{4+t^2}\,dt = \int (1 - \frac{5}{4+t^2})\,dt = t - \frac{5}{2}\operatorname{arctg}\frac{t}{2} + C =$$
$$\cos x - \frac{5}{2}\operatorname{arctg}\frac{\cos x}{2} + C.$$

Case (b). Because both the sine and cosine appear here to an even power, the substitution $t = \operatorname{tg} x$ 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

$$\int \frac{dx}{1+\sin^2 x} = \int \frac{\frac{1}{1+t^2}}{1+\frac{t^2}{1+t^2}}\,dt = \int \frac{1}{1+2t^2}\,dt = \frac{\sqrt{2}}{2}\operatorname{arctg}\left(\sqrt{2}t\right) + C = \frac{\sqrt{2}}{2}\operatorname{arctg}\left(\sqrt{2}\operatorname{tg} x\right) + C.$$

Case (c). Now we'll use the universal substitution $t = \operatorname{tg}\frac{x}{2}$, 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

$$\int \frac{dx}{2-\cos x} = \int \frac{\frac{2}{1+t^2}}{2-\frac{1-t^2}{1+t^2}}\,dt = 2\int \frac{dt}{1+3t^2} = \frac{2\sqrt{3}}{3}\operatorname{arctg}\left(\sqrt{3}t\right) + C = \frac{2\sqrt{3}}{3}\operatorname{arctg}\left(\sqrt{3}\operatorname{tg}\frac{x}{2}\right) + C.$$

□

6.117. Carry out the suggested division of polynomials

$$\frac{2x^5-x^4+3x^2-x+1}{x^2-2x+4}$$

for $x \in \mathbb{R}$. ○

6.118. Express the function

$$y = \frac{3x^4+2x^3-x^2+1}{3x+2}$$

as a sum of a polynomial and a rational function. ○

6.119. 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. ○

6.120. Express the function

$$y = \frac{2x^3+6x^2+3x-6}{x^4-2x^3}$$

in the form of partial fractions. ○

6.121. Decompose the expression

$$\frac{7x^2-10x+37}{x^3-3x^2+9x+13}$$

into partial fractions. ○

6.122. Express the rational function

$$y = \frac{-5x+2}{x^4-x^3+2x^2}$$

in the form of a sum of partial fractions. ○

6.123. Decompose the function

$$y = \frac{1}{x^3(x+1)}$$

into partial fractions. ○

6.124. 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! ○

6.125. 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. ○

6.126. Write the primitive function to the rational function

(a) $y = \frac{3}{x-2}, \quad x \neq 2;$

(b) $y = -\frac{2}{(x-2)^3}, \quad x \neq 2.$

○

6.127. Determine

$$\int \frac{3x+5}{x^2+4x+8}\,dx, \quad x \in \mathbb{R}.$$

○

6.128. Compute the indefinite integral of the function

$$y = \frac{1}{(x^2+x+1)^2}, \quad x \in \mathbb{R}.$$

○

6.129. Determine

$$\int \frac{dx}{x^3+1}, \quad x \neq -1.$$

○

6.130. Integrate

$$\int \frac{1}{x^3-1}\,dx\,, \quad x \neq 1.$$

○

6.131. Compute the integral

$$\int \frac{x^3}{(x-1)(x-2)^2}\,dx\,, \quad x \in \mathbb{R} \setminus \{1, 2\}.$$

○

6.132. 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) $\int \frac{dx}{\sin x}.$

○

6.133. Let $y = |\,x\,|$ on the interval $I = [-1, 1]$ and let

$$\Xi_n = \left(-1, -\tfrac{n-1}{n}, \dots, -\tfrac{1}{n}, 0, \tfrac{1}{n}, \dots, \tfrac{n-1}{n}, 1\right)$$

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).

○

6.134. Compute

$$\lim_{n\to\infty} \frac{\sqrt{1+\frac{1}{n}}+\sqrt{1+\frac{2}{n}}+\cdots+\sqrt{1+1}}{n}.$$

○

6.135. How many distinct primitive functions to function $y = \cos(\ln x)$ does there exist on the interval $(0, 10)$? ○

6.136. Give an example of a function f on th interval $I = [0, 1]$ that doesn't have a primitive function on I. ○

6.137. 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) $\int\limits_{1/\mathrm{e}}^{\mathrm{e}} |\ \ln x\,|\,dx$.

○

6.138. Compute

$$\int\limits_1^2 \frac{x}{\sqrt{1+x^2}}\ dx.$$

○

6.139. For arbitrary real numbers $a < b$ determine

$$\int\limits_a^b \operatorname{sgn} x\ dx.$$

Recall that $\operatorname{sgn} x = 1$, for $x > 0$; $\operatorname{sgn} x = -1$, fpr $x < 0$; and $\operatorname{sgn} 0 = 0$. ○

6.140. Compute the definite integral

$$\int\limits_0^1 \frac{x^3}{1+x^4}\ dx.$$

○

6.141. For example by repeated integration by parts, compute

$$\int\limits_0^{\pi/2} \mathrm{e}^{2x}\ \cos x\ dx.$$

○

6.142. Determine

$$\int\limits_{-1}^1 x^2\ \mathrm{e}^{-x}\,dx\,.$$

○

6.143. Compute the integral

$$\int_{-1}^{1} \frac{x}{\sqrt{5-4x}}\; dx$$

using the substitution method. ○

6.144. Compute

(a) $\int_{1}^{e^8} \frac{dx}{x\sqrt{1+\ln x}}$;

(b) $\int_{0}^{\ln 2} \frac{x}{e^x}\; dx$.

○

6.145. Which of the positive numbers

$$p := \int_{0}^{\pi/2} \cos^7 x\; dx, \quad q := \int_{0}^{\pi} \cos^2 x\; dx$$

is bigger? ○

6.146. Determine the signs of these three numbers (values of integrals)

$$a := \int_{-2}^{2} x^3\, 2^x\; dx; \quad b := \int_{0}^{\pi} \cos x\; dx; \quad c := \int_{0}^{2\pi} \frac{\sin x}{x}\; dx.$$

○

6.147. Order the numbers

$$A := \int_{0}^{\pi/2} \cos x\; \sin^2 x\; dx, \quad B := \int_{0}^{\pi/2} \sin^2 x\; dx, \quad C := \int_{-1}^{1} -x^5\, 5^x\; dx,$$

$$D := \int_{2\pi}^{10} \frac{x^2+2}{x^6+4}\; dx + \int_{\pi}^{2\pi} \frac{x^2+2}{x^6+4}\; dx + \int_{10}^{\pi} \frac{x^2+2}{x^6+4}\; dx$$

by size. ○

6.148. 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) $\int_{0}^{2\pi} \sin x\; dx$.

○

6.149. Compute $\int_{-1}^{1} |\, x \,|\; dx$. ○

6.150. Determine

$$\int_{-1}^{1} x^5\; \sin^2 x\; dx.$$

○

6.151. With an error lesser than 1/10, approximately compute

$$\int_{1}^{2} \left(x - \frac{\cos^{10} x}{10}\right) \ln x\; dx\,.$$

6.152. Without using the symbols of differentiating and integrating, express

$$\left(\int\limits_{x^2}^{a} 3t^2 \cos t \; dt\right)'$$

with variable $x \in \mathbb{R}$ and a real constant a, if we differentiate with respect to x.

6.153. Compute the indefinite integral

$$\int \frac{1}{x^4 + 3x^3 + 5x^2 + 4x + 2}\, \mathrm{d}x.$$

6.154. Compute the integral

$$\int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \frac{\sin t}{1 - \cos^2 t}\, \mathrm{d}t.$$

6.155. Compute the integral

$$\int_0^{\ln 2} \frac{\mathrm{d}x}{e^{2x} - 3e^x}.$$

6.156. Compute:

(i) $\int_0^{\frac{\pi}{2}} \sin x \sin 2x \,\mathrm{d}x$,
(ii) $\int \sin^2 x \sin 2x \,\mathrm{d}x$.

6.157. 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) $\int\limits_{-1}^{1} \ln |\, x \,|\; dx$.

6.158. Determine

$$\int\limits_{0}^{3\pi/2} \frac{\cos x}{1+\sin x}\; dx.$$

6.159. Compute the improper integral

$$\int\limits_{-\infty}^{+\infty} \frac{1}{x^2+x+1}\; dx\,.$$

6.160. Compute

$$\int\limits_{-\infty}^{+\infty} \frac{e^x}{e^{2x}+e^x+1}\, dx\,.$$

○

6.161. By using the substitution method, compute

$$\int\limits_{-\infty}^{0} x\, e^{-x^2}\, dx\,; \quad \int\limits_{0}^{\infty} \frac{e^{-\frac{1}{x}}}{x^2}\, dx\,.$$

○

6.162. Compute the integrals

$$\int\limits_{0}^{1} \frac{e^{-\sqrt{x}}}{\sqrt{x}}\, dx; \quad \int\limits_{1}^{4} \frac{e^{-\sqrt{x}}}{\sqrt{x}}\, dx; \quad \int\limits_{4}^{+\infty} \frac{e^{-\sqrt{x}}}{\sqrt{x}}\, dx.$$

○

6.163. Find the values of $\alpha \in \mathbb{R}$, for which

(a) $\int\limits_{1}^{+\infty} \frac{dx}{x^\alpha} \in \mathbb{R}$;

(b) $\int\limits_{0}^{1} \frac{dx}{x^\alpha} \in \mathbb{R}$;

(c) $\int\limits_{-\infty}^{+\infty} \sin \alpha x\ dx \in \mathbb{R}$.

○

6.164. For which $p, q \in \mathbb{R}$ is the integral

$$\int\limits_{2}^{+\infty} \frac{dx}{x^p \ln^q x}$$

finite? ○

6.165. Decide, if the following is true:

(a) $\int\limits_{-\infty}^{+\infty} \frac{dx}{x^2+3} \in \mathbb{R}$;

(b) $\int\limits_{-\infty}^{+\infty} \frac{dx}{x^2-3} \in \mathbb{R}$;

(c) $\int\limits_{1}^{+\infty} \frac{1+2\sin^3 x}{x^5+x^3+1}\ dx \in \mathbb{R}$.

○

6.166. Approximately compute $\cos \frac{\pi}{10}$ with an error lesser than 10^{-5}. ○

6.167. 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}. ○

6.168. Without computing the derivatives, determine the Talor polynomial of degree 4 centered at $x_0 = 0$ of function

$$f(x) = \cos x - 2\sin x - \ln(1+x)\,, \qquad x \in (-1, 1).$$

Then decide if the graph of function f in neighbourhood of the point $[0, 1]$ is above or below the tangent line. ○

6.169. 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. ○

6.170. 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 $x \in [-1, 1]$. ○

6.171. 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. ○

6.172. 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. ○

6.173. 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. ○

6.174. 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 $a_n x^n$. ○

6.175. Find a function f such that for $x \in \mathbb{R}$, the sequence of functions

$$f_n(x) = \tfrac{n^2x^3}{n^2x^2+1}, \quad n \in \mathbb{N}$$

to it. Is this convergence uniform on $\mathbb{R}$? ○

6.176. Does the series

$$\sum_{n=1}^{\infty} \tfrac{n\,x}{n^4+x^2}, \quad \text{kde} \quad x \in \mathbb{R},$$

converge uniformly on the whole real axis? ○

6.177. 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. ○

6.178. 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}.

○

6.179. Determine the power expansion centered at $x_0 = 0$ of function

$$f(x) = \int\limits_0^x \mathrm{e}^{t^2}\,dt, \quad x \in \mathbb{R}.$$

○

6.180. 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 $x \in [-1, 1]$. ○

6.181. 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. ○

6.182. 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. ○

Solutions of the exercises

6.4. $\frac{2\sin x}{\cos^3 x}$.

6.5. $p^{(5)}(x) = 12 \cdot 5!$; $p^{(6)}(x) = 0$.

6.6. $2^{12}\,\mathrm{e}^{2x} + \cos x$.

6.7. $f^{(26)}(x) = -\sin x + 2^{26}\,\mathrm{e}^{2x}$.

6.14. $\frac{\pi}{4} + \frac{1}{2}(x-1) - \frac{1}{4}(x-1)^2 + \frac{1}{12}(x-1)^3$.

6.15. (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) $x + x^2 + \frac{x^3}{3}$.

6.16. $2(x-1) - (x-1)^2 + \frac{2}{3}(x-1)^3 - \frac{1}{2}(x-1)^4$.

6.17. $\frac{-x^3}{3(1+x)^3}$.

6.18. $x - \frac{x^3}{6}$; $\sin 1^\circ \approx \frac{\pi}{180} - \frac{\pi^3}{6\cdot 180^3}$; $\lim_{x\to 0+} \frac{x\ \sin x - x^2}{x^4} = -\frac{1}{6}$.

6.19. $\sum_{k=0}^{n} \frac{2^k}{k!}\, x^k$, $n \geq 8$, $n \in \mathbb{N}$.

6.20. $(x-1)^3 + 3(x-1)^2 + (x-1) + 4$.

6.26.

$$\sum_{i=0}^{\infty} (-1)^n \frac{2^{2n-1}}{(2n)!} x^{2n},$$

converges for all real x.

6.27.

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{2^{2n-1}}{(2n)!} x^{2n},$$

converges for all real x.

6.28.

$$f(x) = \sum_{n=1}^{\infty} \frac{3(-1)^{n+1}}{n} x^n,$$

converges for $x \in (-1, 1]$.

6.29. 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 $(0, 2]$.

6.31. $\left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$.

6.32. 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 $\pm\infty$).

6.33. (a) $y = 0$ at $-\infty$; (b) $x = 2$ – horizontal, $y = 1$ v $\pm\infty$.

6.34. $y = 0$ for $x \to \pm\infty$.

6.35. $y = \ln 10$, $y = x + \ln 3$.

6.71. $\frac{1}{1-a}$ for $a \in (0, 1)$, ∞ else.

6.89. All of them.

6.90. 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's positive for $x < -1$ and nonpositive for $x > -1$. It can be shown easily that

$$\lim_{x\to -\infty} f(x) = +\infty, \qquad \lim_{x\to +\infty} f(x) = -\infty;$$

$$f'(x) = -\tfrac{x^2+2x}{(x+1)^2}, \qquad f''(x) = -\tfrac{2}{(x+1)^3}, \qquad x \in \mathbb{R} \smallsetminus \{-1\}.$$

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 sharp local minimum $y_2=4$. It's convex on the interval $(-\infty,-1)$ and concave on the interval $(-1,+\infty)$. It doesn't 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, $f(-3)=9/2$, $f'(-3)=-3/4$, $f(1)=-1/2$, $f'(1)=-3/4$.

6.91. The function is defined and continuous on $\mathbb{R}\setminus\{0\}$. It's not odd, even nor periodic. It's negative exactly 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

$$\lim_{x\to 0} f(x)=+\infty,\qquad \lim_{x\to+\infty} f(x)=-\infty,\qquad \lim_{x\to-\infty} f(x)=+\infty.$$

Aditionally,

$$f'(x)=-\tfrac{x^3+2}{x^3},\quad x\in\mathbb{R}\setminus\{0\},$$
$$f''(x)=\tfrac{6}{x^4},\quad x\in\mathbb{R}\setminus\{0\}.$$

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 point x_1 it has a local minimum $y_1=3/\sqrt[3]{4}$. It has no points of inflection. It's convex on its whole domain. The line $x=0$ is a horizontal asymptote and the line $y=-x$ is an inclined asymptote at $\pm\infty$.

6.92. The function is defined and continuous on $\mathbb{R}\setminus\{1\}$. It's 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

$$f(x)=(x-1)^2+\tfrac{2}{x-1},\quad x\in\mathbb{R}\setminus\{1\},$$

it's not difficult to compute

$$f'(x)=2\,\tfrac{(x-1)^3-1}{(x-1)^2},\quad x\in\mathbb{R}\setminus\{1\},$$
$$f''(x)=2\,\tfrac{(x-1)^3+2}{(x-1)^3},\quad x\in\mathbb{R}\setminus\{1\}.$$

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's 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 thus a point of inflection. The line $x=1$ is a horizontal asymptote. The function doesn't have any inclined asymptotes.

6.93. The function is defined and continuous on whole $\mathbb{R}$. It's 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 the point $[0,0]$. The derivative can be determined easily:

$$f'(x)=\tfrac{e^{-x}}{3\sqrt[3]{x^2}}-\sqrt[3]{x}\,e^{-x},\quad x\in\mathbb{R}\setminus\{0\},\qquad f'(0)=+\infty,$$
$$f''(x)=\sqrt[3]{x}\,e^{-x}-\tfrac{2e^{-x}}{3\sqrt[3]{x^2}}-\tfrac{2e^{-x}}{9\sqrt[3]{x^5}},\quad x\in\mathbb{R}\setminus\{0\}.$$

The only zero point of the first derivative is the point $x_0=1/3$. The function f is increasing on the interval $(-\infty,1/3]$ and decreasing on the interval $[1/3,+\infty)$. Hence at the point x_0, it has an absolute maximum $y_0=1/\sqrt[3]{3e}$. Since $\lim_{x\to-\infty} f(x)=-\infty$, its range is $(-\infty,y_0\big]$. 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's 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. $\lim_{x\to+\infty} f(x)=0$.

6.94. The function is defined and continuous on $\mathbb{R}\setminus\{2\}$. It's 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 the point $x_0=2$, the so called jump of size π is realized, as follows from the limits

$$\lim_{x\to 2-} f(x)=\tfrac{\pi}{2},\qquad \lim_{x\to 2+} f(x)=-\tfrac{\pi}{2}.$$

We have

$$f'(x)=\tfrac{1}{x^2-2x+2},\quad x\in\mathbb{R}\setminus\{2\},$$

$$f''(x) = \frac{2\,(1-x)}{(x^2-2x+2)^2}, \quad x \in \mathbb{R} \setminus \{2\}.$$

The first derivative doesn't 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 $(-\pi/2, \pi/2) \setminus \{-\pi/4\}$. 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 $\pm\infty$.

6.95. 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 $y = 0$, $\lim_{x\to0} f(x) = -\infty$, $\lim_{x\to\infty} f(x) = 0$.

6.96. Domain $\mathbb{R} \setminus [1, 2]$. Local maximum $x = \frac{1-\sqrt{5}}{2}$, concave on the whole domain, asymptotes $x = 1$, $x = 2$.

6.97. Domain $\mathbb{R}$. Local minimas -1, 1, maximum at 0. Even function. Points of inflection $\pm\frac{1}{\sqrt{2}}$, no asymptotes.

6.98. Domain $\mathbb{R} \setminus [-\frac{1}{2}, 1]$. No global extremes. No inflection points, asymptotes $x = -\frac{1}{2}$, $x = 1$.

6.99. Domain $\mathbb{R}\setminus\{1\}$. No extrems. No points of inflection, convex on $(-\infty, 1)$, concave on $(1, \infty)$. Horizontal asymptote $x = 1$. Inclined asymptote $y = x + 1$.

6.100. (a) $\frac{8}{15}\, x\sqrt[8]{x^7}$; (b) $\frac{4^x}{\ln 4} + 2\frac{6^x}{\ln 6} + \frac{9^x}{\ln 9}$; (c) $\frac{\arcsin x}{2}$; (d) $\ln(1 + \sin x)$.

6.101. (a) $-\operatorname{cotg} x - x + C$; (b) $\operatorname{tg} x - \operatorname{cotg} x + C$.

6.102. $\mathrm{e}^x + 3\arcsin\frac{x}{2}$.

6.103. $\frac{1}{4}\ln\left(1 + x^4\right) + C$.

6.104. $2\sqrt{2}\operatorname{arctg}\frac{x-1}{\sqrt{2}} + C$.

6.105. $\ln\left|\frac{x^2-1}{x}\right| + \frac{3}{2}\arcsin x - 4\cos x - 5\sin x + C$.

6.106. (a) $x\operatorname{arctg} x - \frac{\ln(1+x^2)}{2} + C$; (b) $\frac{\ln^2 x}{2} + C$.

6.107. (a) $-x^2\cos x + 2x\sin x + 2\cos x + C$; (b) $\mathrm{e}^x\left(x^2 - 2x + 2\right) + C$.

6.108. $\frac{x^2}{4}\left(2\ln^2 x - 2\ln x + 1\right) + C$.

6.109. $\left(2x - x^2\right)\mathrm{e}^x + C$.

6.110. (a) $\frac{(2x+5)^{11}}{22} + C$; (b) $-\frac{1}{\ln x} + C$; (c) $-\frac{1}{3}\,\mathrm{e}^{-x^3} + C$; (d) $5\arcsin^3 x + C$; (e) $\frac{\ln^2 x}{2} + C$; (f) $\operatorname{arctg}^2\sqrt{x} + C$; (g) $\frac{\sqrt{3}}{3}\operatorname{arctg}\left(\frac{\sqrt{3}}{3}\,\mathrm{e}^x\right) + C$; (h) $2\sin\sqrt{x} - 2\sqrt{x}\,\cos\sqrt{x} + C$.

6.111. 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 $\int \frac{2\,dy}{y^2+2y-2}$.

6.112. $\frac{\sqrt{2}}{2}\operatorname{arctg}\left(\sqrt{2}\operatorname{tg} x\right) + C$.

6.113. $x - 2\sqrt{x} + 2\ln\left(1 + \sqrt{x}\right) + C$.

6.114. (a) $\frac{x^{n+1}}{n+1}\ln x - \frac{x^{n+1}}{(n+1)^2} + C$; (b) $\frac{\operatorname{arctg} x^2}{2} + C$.

6.117. $2x^3 + 3x^2 - 2x - 13 + \frac{-19x+53}{x^2-2x+4}$.

6.118. $x^3 - \frac{1}{3}x + \frac{2}{9} + \frac{5}{9(3x+2)}$.

6.119. (a) $\frac{2}{x-2} + \frac{3}{x+2} - \frac{1}{x+3}$; (b) $\frac{2}{x} - \frac{1}{x^2} + \frac{1}{x^2+1} + \frac{x}{(x^2+1)^2}$.

6.120. $\frac{5}{x-2} + \frac{3}{x^3} - \frac{3}{x}$.

6.121. $\frac{3}{x+1} + \frac{4x-2}{x^2-4x+13}$.

6.122. $\frac{1}{x^2} - \frac{2}{x} + \frac{2x-3}{x^2-x+2}$.

6.123. $\frac{1}{x} - \frac{1}{x^2} + \frac{1}{x^3} - \frac{1}{x+1}$.

6.124. $\frac{A}{x-2} + \frac{B}{x^2} + \frac{C}{x} + \frac{Dx+E}{(3x^2+x+4)^2} + \frac{Fx+G}{3x^2+x+4}$.

6.125. $1 + \frac{1}{x^3} - \frac{3}{x} + \frac{5}{x-2}$.

6.126. (a) $3 \ln |\, x - 2\, |$; (b) $\frac{1}{(x-2)^2}$.

6.127. $\frac{3}{2} \ln \left(x^2 + 4x + 8\right) - \frac{1}{2} \operatorname{arctg} \frac{x+2}{2} + C$.

6.128. $\frac{4}{3\sqrt{3}} \operatorname{arctg} \frac{2x+1}{\sqrt{3}} + \frac{2x+1}{3\left(x^2+x+1\right)} + C$.

6.129. $\frac{1}{6} \ln \frac{(x+1)^2}{x^2-x+1} + \frac{\sqrt{3}}{3} \operatorname{arctg} \frac{2x-1}{\sqrt{3}} + C$.

6.130. $\frac{1}{3} \ln |x - 1| - \frac{1}{6} \ln \left(x^2 + x + 1\right) - \frac{1}{\sqrt{3}} \operatorname{arctg} \frac{2x+1}{\sqrt{3}} + C$.

6.131. $\ln \left(|\, x - 1\, |\, (x - 2)^4\right) - \frac{8}{x-2} + x + C$.

6.132. (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) $\ln \left|\operatorname{tg} \frac{x}{2}\right| + C$.

6.133. $S_{\Xi_n,\, \mathrm{sup}} = \frac{n+1}{n}$, $S_{\Xi_n,\, \mathrm{inf}} = \frac{n-1}{n}$; yes, it is.

6.134. $\int_1^2 \sqrt{x}\, dx = \frac{2}{3} \left(2\sqrt{2} - 1\right)$.

6.135. Infinitely many.

6.136. For example, f can attain a value of 1 at rational points of the interval and be zero at irrational points.

6.137. (a) 2; (b) $\frac{\pi}{4} - \frac{\ln 2}{2}$; (c) $2 \ln \left(1 + \sqrt{2}\right)$; (d) $2 - \frac{2}{\mathrm{e}}$.

6.138. $\sqrt{5} - \sqrt{2}$.

6.139. $|b| - |a|$.

6.140. $\frac{1}{4} \ln 2$.

6.141. $\frac{1}{5}\, (\mathrm{e}^{\pi} - 2)$.

6.142. $\mathrm{e} - 5\mathrm{e}^{-1}$.

6.143. $\frac{1}{6}$.

6.144. (a) 4; (b) $\frac{1-\ln 2}{2}$.

6.145. $p < q$.

6.146. $a > 0$; $b = 0$; $c > 0$.

6.147. $C < D = 0 < A < B$.

6.148. (a) 5; (b) 0; (c) 0.

6.149. 1.

6.150. 0.

6.151. $0 < \int_1^2 \frac{\cos^{10} x}{10}\, \ln x\ dx\ < \frac{1}{10}, \int_1^2 x\ \ln x\ dx\ = \ln 4 - \frac{3}{4}$.

6.152. $-6x^5 \cos x^2$.

6.153. $\frac{1}{2} \ln(x^2 + 2x + 2) - \frac{1}{2} \ln(x^2 + x + 1) + \frac{1}{3}\sqrt{3} \arctan \left(\frac{(2x+1)\sqrt{3}}{3}\right) + C$.

6.154. $\frac{1}{2} \ln \left(\frac{2+\ln 2}{2-\ln 2}\right)$.

6.155. $-\frac{1}{6} - \frac{2}{9} \ln 2$.

6.156.

(i) $\frac{2}{3}$,
(ii) $\frac{1}{2} \sin^4 x$.

6.157. (a) π; (b) $+\infty$; (c) 20; (d) -2.

6.158. $-\infty$.

6.159. $\frac{\sqrt{3}}{9}\pi$.

6.160. $\frac{2\sqrt{3}}{9}\pi$.

6.161. $-\frac{1}{2}$; 1.

6.162. $2-\frac{2}{e}$; $\frac{2}{e}-\frac{2}{e^2}$; $\frac{2}{e^2}$.

6.163. (a) $\alpha > 1$; (b) $\alpha < 1$; (c) $\alpha = 0$.

6.164. Exactly for $p > 1$, $q \in \mathbb{R}$ and for $p = 1$, $q > 1$.

6.165. (a) true; (b) false; (c) true.

6.166. $1 - \frac{\pi^2}{10^2 \cdot 2} + \frac{\pi^4}{10^4 \cdot 4!}$.

6.167. The error belongs to the interval (0, 1/200).

6.168. $1 - 3x + \frac{7}{24}x^4$; above the tangent line.

6.169. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$.

6.170. $y = \operatorname{arctg} x$.

6.171. Exactly for $x \in \left(-\frac{5}{2}, \frac{5}{2}\right)$, we have

$$\tfrac{1}{5+2x} = \tfrac{1}{5} \sum_{n=0}^{\infty} \left(-\tfrac{2}{5}\right)^n x^n .$$

6.172. $\frac{1}{3}\sum_{n=0}^{\infty} \frac{2^n}{3^n} x^n$.

6.173.

$$f(x) = 1/2 + \sum_{i=0}^{\infty} \frac{(-1)^{i+1} 2^{2i}}{(2i+1)!} \left(x - \frac{\pi}{4}\right)^{2i+1} .$$

The series converges for all $x \in R$.

6.174. $\sum_{n=0}^{\infty} \frac{e}{n!}(x-1)^n$; $\sum_{n=0}^{\infty} \frac{\ln^n 2}{n!} x^n$.

6.175. $f(x) = x$, $x \in \mathbb{R}$; yes.

6.176. No.

6.177. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$.

6.178. (a) $1 - \frac{\pi^2}{18^2 \cdot 2!} + \frac{\pi^4}{18^4 \cdot 4!}$; (b) $\frac{1}{2} - \frac{1}{5 \cdot 2^5}$.

6.179. $\sum_{n=0}^{\infty} \frac{1}{(2n+1)\, n!} x^{2n+1}$.

6.180. $y = \operatorname{arctg} x$.

6.181. Exactly for $x \in \left(-\frac{5}{2}, \frac{5}{2}\right)$, we have

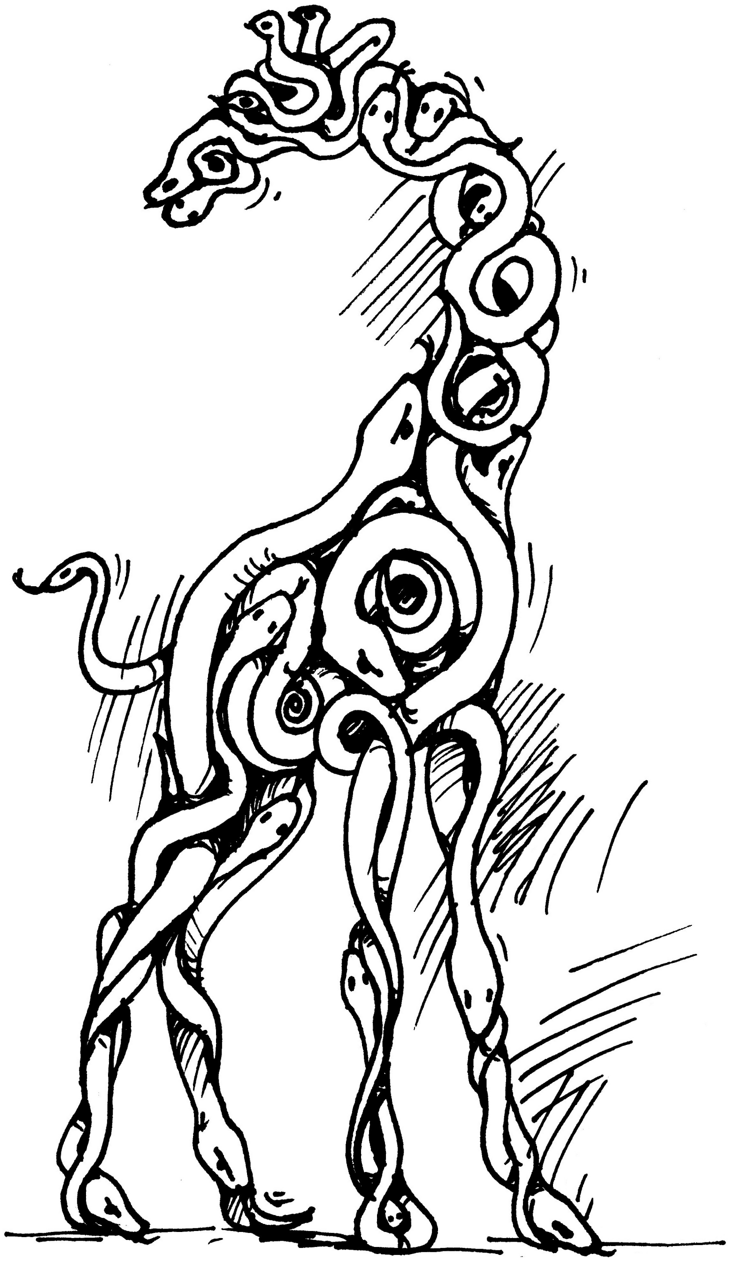
$$\tfrac{1}{5+2x} = \tfrac{1}{5} \sum_{n=0}^{\infty} \left(-\tfrac{2}{5}\right)^n x^n .$$

6.182. (a) $v(0) = 6\,\mathrm{m/s}$; (b) $t = 3\,\mathrm{s}$, $s(3) = 16\,\mathrm{m}$; (c) $v(4) = -2\,\mathrm{m/s}$, $a(4) = -2\,\mathrm{m/s^2}$.

CHAPTER 7

Continuous models

How do we manage non-linear objects?
– mainly by linear tools again...



In this chapter, we will show several applications of the tools of differential and integral calculus for various problems in which we will do with functions of one real variable.

The tools and procedures will be quite similar to the ones shown in the third chapter, i. e. manipulations with linear combinations of selected generators and linear transformations (looking for their kernels or the reverse images of given elements). However, we will work not with finite-dimensional vectors but with spaces of functions, i. e. the vector spaces we will be considering will seldom have finite dimension. We will get back to these as well as other practical fields in the next chapter in the context of functions of several variables, differential equations, and the calculus of variations.

First, we will approximate functions with linear combinations from given sets of generators. However, on the way, we will have to clarify how we can work with such concepts like distance. Actually, we will sketch the basics of what is called the theory of metric spaces, and this part will also serve us as a preparation for analysis in Euclidean spaces $\mathbb{R}^n$. We will mainly resume applying the procedures we have already known from the Euclidean vector spaces. We will find out that our intuition from the Euclidean spaces of low dimensions is quite convenient in general as well.

Then we will focus on integral operators, i. e. linear mappings on functions which are defined in terms of integrals. Especially, we will pay our attention to the so-called Fourier analysis. As usual, our reasoning will touch discrete variants of previously discussed continuous operations.

In the entire chapter, we will work with functions of one variable which will take real or (very often) complex values.

A. Orthogonal systems of functions

If we want to depict a three-dimensional object in a plane, we usually consider its (mostly orthogonal) projection into the plane. Similarly, if we want to "express" some more complicated function in terms of simpler ones, we can consider its projection into the (real) vector space generated by those simpler functions. Then we will be able to integrate, for instance, the more complicated functions in the same way as we integrated (or differentiated) functions expressed in terms of power series (if the space of the simpler functions is "sufficiently" large, then we will be able to do this with arbitrary accuracy).

We can even define a scalar product on a suitable (infinite) vector space of functions on a given interval. Thus the scalar product is not to be defined on the space of all functions on the given interval, but rather on a subspace of its which, on the other hand, will be large enough for our calculations (besides, it will contain all continuous functions on the given interval). The scalar product will allow us to calculate the projections in the same way as we used to do in the case of vector spaces. Given a finite-dimensional vector (sub)space of functions and wanting to determine the projection of a function onto it, we first calculate the orthogonal (or orthonormal) basis of this subspace by the Gram–Schmidt orthogonalization process and then we determine the orthogonal projection in the known way (2.3).

7.1. Let the vector subspace $\langle x^2, 1/x\rangle$ of the space of real-valued functions on the interval $[1, 2]$ be given. Complete the function $1/x$ to an orthogonal basis of the subspace and determine the orthogonal projection of the function x onto it.

Solution. First, we deal with 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 (and all of the vectors lying 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 are looking of 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:

$$\langle \frac{1}{x}, x^2 + k \cdot \frac{1}{x}\rangle = \langle \frac{1}{x}, x^2\rangle + k\langle \frac{1}{x}, \frac{1}{x}\rangle.$$

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 wanted orthogonal basis is $(\frac{1}{x}, x^2 - \frac{3}{x})$. Now, we calculate the projection p_x of the function x onto this subspace (see (2.3)):

$$\begin{aligned} p_x &= \frac{\langle x, \frac{1}{x}\rangle}{\langle \frac{1}{x}, \frac{1}{x}\rangle} \cdot \frac{1}{x} + \frac{\langle x, x^2 - \frac{3}{x}\rangle}{\langle x^2 - \frac{3}{x}, x^2 - \frac{3}{x}\rangle \cdot (x^2 - \frac{3}{x})} \\ &= \frac{2}{x} + \frac{15}{34}(x^2 - \frac{3}{x}). \end{aligned}$$

□

7.2. Let us consider the real vector space of functions on the interval $[1, 2]$ generated by the functions $\frac{1}{x}, \frac{1}{x^2}, \frac{1}{x^3}$. Complete the function $\frac{1}{x}$ to an orthogonal basis of this space.

1. Fourier series

7.1. Spaces of functions. As usual, we begin with choosing appropriate sets of functions which we want to work with. We want to have enough functions so that our models could be conveniently applied in practice. At the same time, the functions must be sufficiently "nice" as we must be able to integrate and differentiate them as needed.

We will largely work with functions defined on an interval $I = [a, b] \subset \mathbb{R}$, or on an infinite interval (i. e. the marginal values a and b can take the values $\pm\infty$, but the sets will still be closed).

Spaces of piecewise smooth functions

The set of functions $\mathcal{S}^0 = \mathcal{S}^0[a, b]$ contains exactly the piecewise continuous functions on $I = [a, b]$ with real or complex values, i. e. we suppose that at every point of the interval, the function $f \in \mathcal{S}^0$ has the corresponding finite one-sided limits both from the left and from the right, and that on every finite interval, there are only finitely many points of discontinuity. Especially, all such functions are bounded on bounded intervals.

For every natural number $k \geq 1$, we will also consider the set of all piecewise continuous functions f such that all their derivatives up to order k (inclusive) lie in $\mathcal{S}^0$ (i. e. they need not exist at all points, but their one-sided limits do exist). We will denote this set by $\mathcal{S}^k$.

In the case of an unbounded interval I, we will also often work with the subset $\mathcal{S}^k_c \subset \mathcal{S}^k$ of all functions with a compact support (i. e. the functions take zero outside some finite closed interval).

On bounded intervals, of course, all such functions have a compact support in this sense. When we are not interested in the interval we work on, we will write only $\mathcal{S}^k_c$ in all cases. In the case of a finite interval $[a, b]$ or under the condition of a compact support, our functions from $\mathcal{S}^0$ are always Riemann integrable on the chosen interval I, both in the absolute value and squared, i. e.

$$\int_a^b |f(x)| dx < \infty, \quad \int_a^b (f(x))^2 dx < \infty.$$

Our reasonings can be extended on significantly greater domains of functions, yet this often costs us a lot of technical effort. From time to time, we will make reference to Kurzweil (or Lebesgue) integrable functions for which the results are more compact and nicer. Interested readers are referred to extended and specialized literature. Actually we will keep the same strategy as with the rational and real numbers – we calculate with nice functions only and we can "handle" the limits of Cauchy sequences in the chosen metrics (which are usually needed only formally).

7.2. Distance of functions. From the already proved properties of limits and derivatives, we can immediately see that $\mathcal{S}^k$ and $\mathcal{S}^k_c$ are vector spaces. In finite-dimensional spaces, the distance of vectors was considered in terms of the differences of the particular coordinates. In spaces of functions, we can proceed analogously and utilize the absolute value of real or complex numbers (or the Euclidean distance) in the following way:

Distance of functions

Solution. Similarly to the previous exercise, we use the Gram–Schmidt orthogonalization process (with the given scalar product). Thus we gradually get that
$$\begin{aligned} f_1(x) &= \tfrac{1}{x}, \\ f_2(x) &= \tfrac{1}{x^2} - \tfrac{3}{4x}, \\ f_3(x) &= \tfrac{1}{x^3} - \tfrac{3}{2x^2} + \tfrac{13}{24x}. \end{aligned}$$
□

7.3. Determine the projection of the functions $\frac{1}{x^4}$ and x onto the vector space from the exercise ‖7.2‖. Determine the distances from this vector space as well.

Solution. Projection of $\frac{1}{x^4}$: $\frac{15}{32}f_1+\frac{69}{40}f_2+\frac{9}{4}f_3$, distance: $\frac{\sqrt{14}}{2240}$. Projection of x : $2f_1 + (-\frac{3}{4} + \ln(2))f_2 + (-\frac{3}{2}\ln(2) + \frac{25}{24})f_3$, distance: about 0.03496029493. We can see that the distance of the function which behaves in a similar way as the generators is smaller.
□

7.4. Let the vector subspace $\langle \sin(x), x\rangle$ of the space of real-valued functions on the interval $[0, \pi]$ be given. 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. ○

7.5. Let the vector subspace $\langle \cos(x), x\rangle$ of the space of real-valued functions on the interval $[0, \pi]$ be given. Complete the function $\cos(x)$ to an orthogonal basis of the subspace and determine the orthogonal projection of the function $\frac{1}{3}\cos(x)$ onto it. ○

B. Fourier series

One of the fundamental studied periodic processes which can be met in applications is a general simple harmonic oscillation in mechanics. It is the case of a mass point moving along a straight line. It is well known that the function f which describes the position of the mass point on the line in time is of the form

$$f(t) = a\sin(\omega t + b) \tag{7.1}$$

for certain constants $a, \omega > 0$, $b \in \mathbb{R}$ determined by the position and velocity of the point at the initial time. The function $y \equiv f(t)$ can, for instance, be obtained by solving the homogeneous linear differential equation

$$y'' + \omega^2 y = 0 \tag{7.2}$$

following from Newton's law of force for the given movement. Let us mention that the function f has period $T = 2\pi/\omega$ (in mechanics, one often talks about frequency $1/T$) and that the positive value a (expressing the maximum displacement of the oscillating point from the equilibrium position) is called the amplitude. The value b (expressing the position of the point at the initial time) is called the initial phase, and ω is the angular frequency of the oscillation.

Similarly, we can focus on the function $z \equiv g(t)$ which 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

$$z'' + \omega^2 z = 0. \tag{7.3}$$

The only difference between the equations (‖7.2‖) ans (‖7.3‖) (besides the dissimilar physical interpretation) is the constant ω. In the equation (‖7.2‖), there is $\omega^2 = k/m$ where k is the proportionality constant

Definition. The L_1–distance of functions f and g from $\mathcal{S}^0_c$ is defined by

$$\|f - g\|_1 = \int_a^b |f(x) - g(x)|dx\,.$$

Similarly, L_2–distance of f and g is defined by

$$\|f - g\|_2 = \left(\int_a^b |f(x) - g(x)|^2 dx\right)^{1/2}.$$

The size of the function $\|f\|_1$ or $\|f\|_2$ is understood to be its distance from the zero function.

In the first case, the L_1–distance of functions f and g which take real values only expresses the area enclosed by the graphs of these functions, regardless of which function takes greater values. Since we consider piecewise continuous functions f and g, their distance can equal zero only if they differ in their values at the points of discontinuity, i. e. at only finitely many points on bounded intervals. Indeed, if our functions differ at a point x_0 and they are continuous at it, then they also differ on some sufficiently small neighborhood of this point, and this neighborhood, in turn, contributes a non-zero value into the integral.

If we have three functions f, g, and h, then, of course,

$$\begin{aligned} \int_a^b |h(x) - f(x)|dx &= \int_a^b |h(x) - g(x) + g(x) - f(x)|dx \\ &\le \int_a^b |h(x) - g(x)|dx + \int_a^b |g(x) - f(x)|dx, \end{aligned}$$

so the usual triangle inequality holds. We can notice that to derive this inequality, we only used the triangle inequality for the scalars; thus it holds for functions $f, g \in \mathcal{S}^0_c$ with complex values as well.

The second definition is similar. The square of the size $\|f\|_2$ of a function f is

$$\|f\|^2{}_2 = \int_a^b |f(x)|^2\, dx$$

and it is derived from the well-defined symmetric bilinear mapping of real functions to scalars

$$\langle f, g\rangle = \int_a^b f(x)g(x)\, dx$$

by substituting f for both functions. In the case of complex values, we obtain this size similarly from the scalar product, using complex conjugation:

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}\, dx,$$

as we saw when talking about the unitary spaces in the third chapter:

Thus the triangle inequality will hold as well because the whole discussion can be done in a space of dimension at most three with scalar product, generated by given functions f, g, h.

7.3. (In)finiteness of dimensions, orthogonality. Let us, for a while, stay at our definition of the L_2–size $\|\ \|_2$ on the vector space $\mathcal{S}^0_c$. Apparently, the operation at the end of the last paragraph satisfies both linearity in the first argument and symmetry

$$\langle f, g\rangle = \overline{\langle g, f\rangle},$$

and m is the mass of the point, while in the equation (‖7.3‖), there is $\omega^2 = (LC)^{-1}$.

Actually, every periodic process which can be described by a function of the form (‖7.1‖) is considered a harmonic oscillation, and the constants a, ω, b are almost exclusively given the mentioned names borrowed from the simple harmonic oscillation of a mass point in mechanics.

Applying one of the sum formulae

$$\sin(\alpha+\beta) = \cos\alpha\sin\beta + \sin\alpha\cos\beta, \qquad \alpha, \beta \in \mathbb{R},$$

we can write the function f (see (‖7.1‖)) as

$$f(t) = c\cos(\omega t) + d\sin(\omega t)\,, \tag{7.4}$$

where $c = a\sin b$, $d = a\cos b$. Thus, the function f from (‖7.4‖) also describes a harmonic oscillation with amplitude $a = \sqrt{c^2+d^2}$ and the initial phase $b \in [0, 2\pi)$ satisfying $\sin b = c/a$, $\cos b = d/a$.

An important task in application problems is the composition (so-called superposition) of different harmonic oscillations. A key position is occupied by the superposition of finitely many harmonic oscillations expressed by functions of the form

$$f_n(x) = a_n\cos(n\omega x) + b_n\sin(n\omega x)$$

for $n \in \{1, \dots, m\}$. These particular functions have prime period $2\pi/(n\omega)$. Therefore, their sum

$$\sum_{n=1}^{m} [a_n\cos(n\omega x) + b_n\sin(n\omega x)] \tag{7.5}$$

is a periodic function with period $2\pi/\omega$. It holds generally that the superposition of any finite collection of simple harmonic oscillations with commensurable periods is a period function whose period is the lease common multiple of the prime periods of the particular oscillations.

The sum (‖7.5‖) modified by an appropriate movement,

$$\frac{a_0}{2} + \sum_{n=1}^{m} [a_n\cos(n\omega x) + b_n\sin(n\omega x)]\,, \tag{7.6}$$

is just the m-th partial sum of the functional series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n\cos(n\omega x) + b_n\sin(n\omega x)]\,. \tag{7.7}$$

From the physical point of view, it is a complicated periodic process which can serve as a natural approximation of the superposition of infinitely many simple harmonic oscillations (so-called harmonic components) of the functional series (‖7.7‖).

It may be an interesting question whether, on the other hand, every periodic process can be in a "reasonable" way expressed by the superposition of finitely (or possibly infinitely) many simple harmonic oscillations – whether every periodic process is the result of such a superposition. Exactly formulated from the mathematical point of view: whether every periodic function can be expressed as the finite sum (‖7.6‖), or at least as the series (‖7.7‖). Of course, the positive answer for a significant and broad class of periodic functions is obtained for the infinite sum only (see the theoretical part).

We have already mentioned that periodic processes play an important role in many physical and technical fields. Traditionally, we i. e. in the real case, it is a symmetric bilinear mapping. At the same time, for continuous functions, the condition of non-zero size of non-zero functions holds as well, while for our piecewise continuous functions, zero size implies that the function is non-zero except for an (at most) countable set of points (finite on every finite interval). Thus we have truly defined the scalar product for the vector subspace of continuous functions.

In the case of more general functions, we should, from the technical point of view, identify functions which differ on finite intervals at finitely many points only. In our subsequent reasonings, this technical obstruction will play an insignificant role (we will occasionally make a reference to it in notes).

In the case of finite-dimensional real or complex vector spaces, we considered scalar products and the size of vectors as soon as in the second and third chapter. Now let us notice that when we derived the properties, we always worked with pairs or finite sets of vectors.

Now, we can do just the same with functions, and if we restrict our definition of a scalar product to a vector subspace generated (over real or complex numbers, according to our need) by only finitely many functions $f_1, \dots, f_k$, we again obtain a well-defined scalar product on this finite-dimensional vector subspace.

As an example, let us consider the functions $f_i = x^i$, $i = 0, \dots, k$. In $\mathcal{S}^0$, they generate the $(k+1)$–dimensional vector subspace $\mathbb{R}_k[x]$ of all polynomials of degree at most k. The scalar product of two such polynomials is given by an integral. Every polynomial of degree at most k is uniquely expressed as a linear combination of the generators $f_0, \dots, f_k$. Moreover, if our generators were such that

$$\langle f_i, f_j\rangle = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases} \tag{7.1}$$

then we would have the so-called *orthonormal basis*. At this occasion, let us remind the procedure of Gram–Schmidt orthogonalization, see 2.42, which 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 as $g_1 = f_1$ and by the formulae

$$g_{\ell+1} = f_{\ell+1} + a_1 g_1 + \dots + a_\ell g_\ell, \quad a_i = -\frac{\langle f_{\ell+1}, g_i\rangle}{\|g_i\|^2}$$

for $\ell \geq 1$.

For illustration, we will apply this procedure to three polynomials $1, x, x^2$ on the interval $[-1, 1]$. We get $g_1 = 1$,

$$g_2 = x - \frac{1}{\|g_1\|^2}\int_{-1}^{1} x\cdot 1\,dx \cdot 1 = x - 0 = x$$

$$g_3 = x^2 - \frac{1}{\|g_1\|^2}\int_{-1}^{1} x^2\cdot 1\,dx\cdot 1 - \frac{1}{\|g_2\|^2}\int_{-1}^{1} x^2\cdot x\,dx\cdot x$$
$$= x^2 - \frac{1}{3}.$$

Thus 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 $1, x, x^2 - 1/3$. Normalization, i. e. multiplying by an appropriate scalar to change the size of the basis' elements to one, gives the orthonormal basis

$$h_1 = \sqrt{\frac{1}{2}}, \quad h_2 = \sqrt{\frac{3}{2}}x, \quad h_3 = \frac{1}{2}\sqrt{\frac{5}{2}}(3x^2 - 1/3).$$

can point out acoustics, mechanics, and electrical engineering where answering the above question is undoubtedly of extreme importance. Besides that, looking for the answer has given rise to an independent mathematical field – the theory of Fourier series. Later, it began to be applied in many other classes of problems (among others, for solving the most of important types of ordinary and partial differential equations) and led to development of the particular theoretical foundations of mathematics (for instance, to precise definition of the fundamental concepts of a function and an integral).

The Fourier series are named in honor of French mathematician and physicist Jean B. J. Fourier, who was the first to apply trigonometric expressions (‖7.6‖) in practice in his work from 1822 devoted to the issue of heat conduction (he began to deal with this issue in 1804, and he finished the publication as early as in 1811). The significance of this work of Fourier's for physics is enormous although Fourier himself did not pay much attention to physics. He introduced mathematical methods which even nowadays belong to the standard tools of theoretical physics. His mathematical theory of heat also became the foundations for George S. Ohm when he derived the famous law of conduction of electric current. We should not forget to mention that there were many mathematicians who studied the properties of the sums (‖7.6‖) many years earlier than Fourier (L. Euler, for example). However, they did not achieve such significant results with regard to practical applications.

7.6. 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. The case (a). Since for $x \in \mathbb{R}$, we have

$$\sin(2x)\cos(3x) = \sin(2x)\left[\cos(2x)\cos x - \sin(2x)\sin x\right] = \tfrac{1}{2}\sin(4x)\cos x - \sin^2(2x)\sin x = \tfrac{1}{2}\cos x\sin(4x) - \tfrac{1-\cos(4x)}{2}\sin x = -\tfrac{1}{2}\sin x + \tfrac{1}{2}\cos x\sin(4x) + \tfrac{1}{2}\sin x\cos(4x) = -\tfrac{1}{2}\sin x + \tfrac{1}{2}\sin(5x)\ ,$$

we can see that the Fourier coefficients are all zero except for $b_1 = -1/2, b_5 = 1/2$.

The case (b). Similarly, from

$$\cos^4 x = \left[\cos^2 x\right]^2 = \left[\tfrac{1+\cos(2x)}{2}\right]^2$$
$$= \tfrac{1}{4}\left[1 + 2\cos(2x) + \cos^2(2x)\right] = \tfrac{1}{4}\left[1 + 2\cos(2x) + \tfrac{1+\cos(4x)}{2}\right] = \tfrac{3}{8} + \tfrac{1}{2}\cos(2x) + \tfrac{1}{8}\cos(4x)\ , \quad x \in \mathbb{R},$$

it follows that $a_0 = 3/4, a_2 = 1/2, a_4 = 1/8$, and the other coefficients are all zero.

We showed in this exercise that the calculation of the Fourier series may not lead to integrations (usually by parts). Especially in the cases where the function g is a product (power) of functions $y = \sin(mx)$, $y = \cos(nx)$ for $m, n \in \mathbb{N}$, it suffices to apply high-school knowledge (well-known trigonometric formulae). □

7.7. 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) $g(x) = 0,\ x \in [-1, 0), \quad g(x) = x + 1,\ x \in [0, 1)$.

Such orthonormal generators of $\mathbb{R}_k[x]$ are called *Legendre polynomials.*

7.4. Orthogonal systems of functions. We have just reminded

ourselves the advantages of orthonormal bases of subspaces of finite-dimensional vector spaces. In the last example of Legendre polynomials generating $\mathbb{R}_2[x] \subset V = \mathbb{R}_k[x]$, $k \geq 2$, for any polynomial $h \in V$, the function

$$H = \langle h, h_1\rangle h_1 + \langle h, h_2\rangle h_2 + \langle h, h_3\rangle h_3$$

will be the uniquely determined function which minimizes our L_2–distance $\|h - H\|$ among all functions in $\mathbb{R}_k[x]$, see 3.25.

The coefficients for the best approximation of a given function by a function from a selected subspace can be obtained just by integrating in the definition of the scalar product.

The mentioned example suggests the following generalization: If we do the Gram–Schmidt orthogonalization for all monomials $1, x, x^2, \ldots$, i. e. for a countable system of generators, what will become of that?

Orthogonal systems of functions

Every (at most) countable system of linearly independent functions in $\mathcal{S}^0_c[a, b]$ such that the scalar 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 for all n, the size $\|f_n\|_2 = 1$, we talk about an *orthonormal system of functions.*

Let us thus 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

$$F(x) = \sum_{n=1}^{\infty} c_n f_n$$

converges uniformly on a finite interval $[a, b]$. Then the scalar product $\langle F, f_n\rangle$ can easily be expressed in terms of the particular summands (see the corollary 6.43), obtaining

$$\langle F, f_n\rangle = \sum_{m=1}^{\infty} c_m \int_a^b f_m(x)\overline{f_n(x)}\, dx = c_n\|f_n\|^2,$$

where the norm means (just like in further paragraphs) our L_2–size.

Surely we can now anticipate in what sense the procedures

from finite-dimensional spaces can be extended: Instead of finite linear combinations of base vectors, we will work with infinite series of pairwise orthogonal functions. The following theorem gives us a transparent and very general answer to the question how well the finite sums of such a series can approach a given function:

7.5. 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. Let us denote*

$$c_n = \|f_n\|^{-2}\int_a^b g(x)\overline{f_n(x)}\, dx.$$

(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

Solution. The case (a). Direct calculation gives

$$\begin{aligned}
a_0 &= \tfrac{1}{\pi}\int\limits_{x_0}^{x_0+2\pi} g(x)\,dx = \tfrac{1}{\pi}\int\limits_{-\pi}^{0} 0\,dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi}\sin x\,dx\\
&= \tfrac{1}{\pi}\left[-\cos x\right]_0^{\pi} = \tfrac{2}{\pi},\\
a_n &= \tfrac{1}{\pi}\int\limits_{x_0}^{x_0+2\pi} g(x)\cos(nx)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{-\pi}^{0} 0\,dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi}\sin x\cos(nx)\ dx\\
&= \tfrac{1}{2\pi}\int\limits_{0}^{\pi}\sin([1+n]x) + \sin([1-n]x)\ dx\\
&= \tfrac{1}{2\pi}\left[-\tfrac{\cos([1+n]x)}{1+n} - \tfrac{\cos([1-n]x)}{1-n}\right]_0^{\pi}\\
&= \tfrac{1}{2\pi}\left(-\tfrac{\cos([1+n]\pi)}{1+n} - \tfrac{\cos([1-n]\pi)}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n}\right), \quad n\in\mathbb{N},\\
b_1 &= \tfrac{1}{\pi}\int\limits_{x_0}^{x_0+2\pi} g(x)\sin x\,dx = \tfrac{1}{\pi}\int\limits_{-\pi}^{0} 0\,dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi}\sin^2 x\,dx\\
&= \tfrac{1}{2\pi}\int\limits_{0}^{\pi} 1-\cos(2x)\ dx = \tfrac{1}{2\pi}\left[x - \tfrac{\sin(2x)}{2}\right]_0^{\pi} = \tfrac{1}{2},\\
b_n &= \tfrac{1}{\pi}\int\limits_{x_0}^{x_0+2\pi} g(x)\sin(nx)\ dx = \tfrac{1}{\pi}\int\limits_{-\pi}^{0} 0\,dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi}\sin x\sin(nx)\ dx\\
&= \tfrac{1}{2\pi}\int\limits_{0}^{\pi}\cos([1-n]x) - \cos([1+n]x)\ dx\\
&= \tfrac{1}{2\pi}\left[\tfrac{\sin([1-n]x)}{1-n} - \tfrac{\sin([1+n]x)}{1+n}\right]_0^{\pi} = 0,\ \text{for } n\in\mathbb{N}\smallsetminus\{1\}
\end{aligned}$$

Thus, we get the Fourier series

$$\tfrac{1}{\pi} + \tfrac{\sin x}{2} + \tfrac{1}{2\pi}\sum_{n=1}^{\infty}\left[\left(-\tfrac{\cos([1+n]\pi)}{1+n} - \tfrac{\cos([1-n]\pi)}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n}\right)\cos(nx)\right].$$

The obtained result can be refined: For even numbers n, we have

$$\begin{aligned}
&-\tfrac{\cos([1+n]\pi)}{1+n} - \tfrac{\cos([1-n]\pi)}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n}\\
&\qquad = \tfrac{1}{1+n} + \tfrac{1}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n} = -\tfrac{4}{n^2-1},
\end{aligned}$$

and for odd numbers n,

$$\begin{aligned}
&-\tfrac{\cos([1+n]\pi)}{1+n} - \tfrac{\cos([1-n]\pi)}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n}\\
&\qquad = -\tfrac{1}{1+n} - \tfrac{1}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n} = 0.
\end{aligned}$$

Altogether,

$$-\tfrac{\cos([1+n]\pi)}{1+n} - \tfrac{\cos([1-n]\pi)}{1-n} + \tfrac{1}{1+n} + \tfrac{1}{1-n} = 2\,\tfrac{(-1)^{n+1}-1}{n^2-1}, \quad n\in\mathbb{N},$$

so the resulting series can be written as

$$\tfrac{1}{\pi} + \tfrac{\sin x}{2} + \tfrac{1}{\pi}\sum_{n=1}^{\infty}\left[\tfrac{(-1)^{n+1}-1}{n^2-1}\cos(nx)\right] = \tfrac{1}{\pi} + \tfrac{\sin x}{2} - \tfrac{2}{\pi}\sum_{n=1}^{\infty}\tfrac{\cos(2nx)}{4n^2-1}.$$

The case (b). First, let us mention that the given function is often talked of as a function of sawtooth-shaped oscillation and that its expression as a Fourier series is very important in practice. Since the function g is even on $(-\pi,\pi)$, we can immediately see that $b_n = 0$ for all $n\in\mathbb{N}$. Therefore, it suffices to determine

$$a_0 = \tfrac{1}{\pi}\int\limits_{x_0}^{x_0+2\pi} g(x)\,dx = \tfrac{2}{\pi}\int\limits_{0}^{\pi} x\,dx = \tfrac{2}{\pi}\left[\tfrac{x^2}{2}\right]_0^{\pi} = \pi,$$

$f_1,\dots,f_n$ is

$$h_n = \sum_{i=1}^{n} c_i f_i(x).$$

(2) The series $\sum_{n=1}^{\infty}|c_n|^2\|f_n\|^2$ always converges and it holds that

$$\sum_{n=1}^{\infty}|c_n|^2\|f_n\|^2 \le \|g\|^2.$$

(3) L_2–distance of g from the partial sums $s_k = \sum_{n=1}^{k} c_n f_n$ converges to zero, i. e.

$$\lim_{k\to\infty}\|g - s_k\| = 0,$$

if and only if

$$\sum_{n=1}^{\infty} c_n^2\|f_n\|^2 = \|g\|^2.$$

Before we start with the proof, let us first look at the meanings of the particular 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, we will surely be able to approximate even functions only. Nevertheless, the first statement of the theorem says that we can always reach the best approximation possible by partial sums (in L_2–distance).

The second and third statements then can be perceived as an analogy to the orthogonal projections into subspaces expressed by Cartesian coordinates. Indeed, if for a given function g, the series $F(x) = \sum_{n=1}^{\infty} c_n f_n$ converges pointwise, then the function $F(x)$ is, in a certain sense, a orthogonal projection of g into the vector subspace of all such series.

The second statement is called *Bessel's inequality* and it is an analogy of the finite-dimensional proposition that the orthogonal projection of a vector cannot be greater 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 become smaller by the orthogonal projection into a given subspace, then it surely belongs to the subspace.

On the other hand, our theorem does not claim that the partial sums of the considered series would have 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 (i. e. if we considered more general functions than the ones from our space $\mathcal{S}^0[a,b]$) even in the case when the equality in (3) holds. However, if the series $\sum_{n=1}^{\infty}|c_n|$ converges to a finite value and all the functions f_n are bounded uniformly on I, then, the series $F(x) = \sum_{n=1}^{\infty} c_n f_n$ apparently converges at every point x. Yet it need not converge to the function g everywhere. We will get back to this problem shortly.

The proof of all of the three statements of the theorem is quite similar to the case of finite-dimensional Euclidean spaces. No wonder it is so as the bounds for the distances of g from the partial sum f are constructed in the finite-dimensional linear hull of the functions concerned:

and for any $n \in \mathbb{N}$ using integration by parts, we get

$$\begin{aligned} a_n &= \tfrac{1}{\pi}\int_{x_0}^{x_0+2\pi} g(x)\cos(nx)\ dx = \tfrac{2}{\pi}\int_0^{\pi} x\cos(nx)\ dx \\ &= \tfrac{2}{\pi}\left[\tfrac{x}{n}\ \sin(nx)\right]_0^{\pi} - \tfrac{2}{n\pi}\int_0^{\pi}\sin(nx)\ dx \\ &= \tfrac{2}{n^2\pi}\left[\cos(nx)\right]_0^{\pi} = \tfrac{2}{n^2\pi}\left[(-1)^n - 1\right], \text{ so} \end{aligned}$$

$$a_n = -\tfrac{4}{n^2\pi} \text{ for } n \text{ odd}, \quad a_n = 0 \text{ for } n \text{ even}.$$

Now, we know the Fourier series of the function of sawtooth-shaped oscillation

$$\tfrac{\pi}{2} + \tfrac{2}{\pi}\sum_{n=1}^{\infty}\left[\tfrac{(-1)^n-1}{n^2}\cos(nx)\right] = \tfrac{\pi}{2} - \tfrac{4}{\pi}\sum_{n=1}^{\infty}\tfrac{\cos([2n-1]x)}{(2n-1)^2} =$$
$$\tfrac{\pi}{2} - \tfrac{4}{\pi}\left[\cos x + \tfrac{\cos(3x)}{3^2} + \tfrac{\cos(5x)}{5^2} + \cdots\right].$$

This series could have been found by an easier means – integrating the Fourier series of Heaviside's function (see "angular wave function" in the theoretical part).

The case (c). The function's period is $T = 2$, so we use more general formulae

$$\begin{aligned} a_0 &= \tfrac{2}{T}\int_{x_0}^{x_0+T} g(x)\,dx = \int_{-1}^{1} g(x)\,dx = \int_{-1}^{0} 0\,dx + \int_{0}^{1}(x+1)\,dx = \tfrac{3}{2}, \\ a_n &= \tfrac{2}{T}\int_{x_0}^{x_0+T} g(x)\cos(n\omega x)\ dx = \int_{-1}^{1} g(x)\cos(n\pi x)\ dx \\ &= \int_{-1}^{0} 0\,dx + \int_{0}^{1}(x+1)\cos(n\pi x)\ dx = \tfrac{(-1)^n-1}{n^2\pi^2}, \quad n \in \mathbb{N}, \\ b_n &= \tfrac{2}{T}\int_{x_0}^{x_0+T} g(x)\sin(n\omega x)\ dx = \int_{-1}^{1} g(x)\sin(n\pi x)\ dx \\ &= \int_{-1}^{0} 0\,dx + \int_{0}^{1}(x+1)\sin(n\pi x)\ dx = \tfrac{1-2(-1)^n}{n\pi}, \quad n \in \mathbb{N}. \end{aligned}$$

The calculation of a_0 was simple and needs no more comments. As for determining the integrals at a_n and b_n, it again sufficed to use integration by parts once (differentiating the polynomial $u = x + 1$). Thus, the wanted Fourier series is

$$\tfrac{3}{4} + \sum_{n=1}^{\infty}\left(\tfrac{(-1)^n-1}{n^2\pi^2}\cos(n\pi x) + \tfrac{1-2(-1)^n}{n\pi}\sin(n\pi x)\right).$$

Some refinements of the expression can be achieved when we realize, for instance, that for $n \in \mathbb{N}$, we have

$$a_n = -\tfrac{2}{n^2\pi^2} \text{ for } n \text{ odd}, \quad a_n = 0 \text{ for } n \text{ even},$$

and, similarly,

$$b_n = \tfrac{3}{n\pi} \text{ for } n \text{ odd}, \quad b_n = -\tfrac{1}{n\pi} \text{ for } n \text{ even}.$$

□

7.8. Let the Fourier series of a function f on the interval $[-\pi, \pi]$ with coefficients a_m, b_n, $m \in \mathbb{N} \cup \{0\}$, $n \in \mathbb{N}$ be given. 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 the calculations

Proof of theorem 7.5. Let us choose a linear combination $f = \sum_{n=1}^{k} a_n f_n$ and calculate its distance from g. We get

$$\begin{aligned} \|g - \sum_{n=1}^{k} a_n f_n\|^2 &= \int_a^b \left|g(x) - \sum_{n=1}^{k} a_n f_n(x)\right|^2 dx \\ &= \int_a^b |g(x)|^2\,dx - \int_a^b \sum_{n=1}^{k} g(x)\overline{a_n f_n(x)}\,dx - \\ &\quad - \int_a^b \sum_{n=1}^{k} a_n f_n(x)\overline{g(x)}\,dx + \int_a^b \left|\sum_{n=1}^{k} a_n f_n(x)\right|^2 dx \\ &= \|g\|^2 - \sum_{n=1}^{k}\overline{a_n}c_n\|f_n\|^2 - \sum_{n=1}^{k} a_n\overline{c_n}\|f_n\|^2 + \sum_{n=1}^{k} a_n^2\|f_n\|^2 \\ &= \|g\|^2 + \sum_{n=1}^{k}\|f_n\|^2\big((c_n - a_n)\overline{(c_n - a_n)} - |c_n|^2\big). \end{aligned}$$

Apparently, the last expression can be minimized be choosing $a_n = c_n$, which finishes the proof of the first statement.

Substituting this choice yields the so-called *Bessel's identity*

$$\|g - \sum_{n=1}^{k} c_n f_n\|^2 = \|g\|^2 - \sum_{n=1}^{k}|c_n|^2\|f_n\|^2,$$

from which the Bessel's inequality

$$\sum_{n=1}^{k} c_n^2\|f_n\|^2 \le \|g\|^2$$

immediately follows as the left-hand side is non-negative. Therefore, the whole second statement has been proved as well because every non-decreasing sequence of real numbers which is bounded from above has a limit (the limit is equal to the supremum of the set of the sequence's terms).

If the Bessel's inequality happens to hold with equality, then the statement (3) follows straight from the definitions and the Bessel's identity proved above. □

An orthogonal system of functions is called a *complete orthogonal system* on an interval $I = [a, b]$ for some space of functions on I iff Parseval's equality holds for every function g from this space.

7.6. 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 say what the whole space 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.

$$
\begin{aligned}
a_{2k-1} &= \tfrac{1}{\pi}\int\limits_{-\pi}^{\pi} f(x)\cos([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{-\pi}^{0} f(x)\cos([2k-1]x)\ dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k-1]x)\ dx\\
&= |\, x = y+\pi \,| = \tfrac{1}{\pi}\int\limits_{-2\pi}^{-\pi} f(y+\pi)\cos([2k-1][y+\pi])\ dy\\
&\quad + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\cos([2k-1][y+\pi])\ dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\big[\cos([2k-1]y)\cos([2k-1]\pi)\\
&\quad - \sin([2k-1]y)\sin([2k-1]\pi)\,\big]dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k-1]x)\ dx\\
&= -\tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\cos([2k-1]y)\ dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k-1]x)\ dx = 0,\\
b_{2k-1} &= \tfrac{1}{\pi}\int\limits_{-\pi}^{\pi} f(x)\sin([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{-\pi}^{0} f(x)\sin([2k-1]x)\ dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\sin([2k-1]x)\ dx\\
&= |\, x = y+\pi \,| = \tfrac{1}{\pi}\int\limits_{-2\pi}^{-\pi} f(y+\pi)\sin([2k-1][y+\pi])\ dy\\
&\quad + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\sin([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\sin([2k-1][y+\pi]) + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\sin([2k-1]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\big[\sin([2k-1]y)\cos([2k-1]\pi) + \sin([2k-1]\pi)\cos([2k-1]y)\big]\ dy\\
&\quad + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\sin([2k-1]x)\ dx\\
&= -\tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\sin([2k-1]y)\ dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\sin([2k-1]x)\ dx = 0.
\end{aligned}
$$

The case (b). We immediately get

$$a_0 = \tfrac{1}{\pi}\int\limits_{-\pi}^{\pi} f(x)\,dx = \tfrac{1}{\pi}\int\limits_{-\pi}^{0} f(x)\,dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\,dx = 0,$$

and then, in an analogous way as for the first statement, we get that for any $k \in \mathbb{N}$,

$$
\begin{aligned}
a_{2k} &= \tfrac{1}{\pi}\int\limits_{-\pi}^{\pi} f(x)\cos([2k]x)\ dx\\
&= \tfrac{1}{\pi}\int\limits_{-\pi}^{0} f(x)\cos([2k]x)\ dx + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k]x)\ dx = |\, x = y+\pi \,|\\
&= \tfrac{1}{\pi}\int\limits_{-2\pi}^{-\pi} f(y+\pi)\cos([2k][y+\pi])\ dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k]x)\ dx\\
&= -\tfrac{1}{\pi}\int\limits_{0}^{\pi} f(y)\cos([2k][y+\pi])\ dy + \tfrac{1}{\pi}\int\limits_{0}^{\pi} f(x)\cos([2k]x)\ dx
\end{aligned}
$$

We talk about (abstract) *Fourier series* and the coefficients c_n from the previous theorem are called *Fourier coefficients* of a given function.

In the case when we have an orthonormal system of functions f_n, the formulae mentioned in the theorem are a bit simpler, but still there is no further improvement.

The choice of an orthogonal system of functions for use in practice must follow the purpose for which we want to apply the approximation and further tools. The name "Fourier series" itself references to the following choice of a system of real-valued functions:

Fourier's orthogonal system

$$1,\ \sin x,\ \cos x,\ \sin 2x,\ \cos 2x,\ \dots,\ \sin nx,\ \cos nx,\ \dots$$

As an elementary exercise on integration by parts, we can calculate that indeed it is an orthogonal system of functions on the interval $[-\pi, \pi]$. Presently, we will show another means of verification of this fact.

These functions are periodic with common period 2π (see the definition below) and the so-called "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 so-called signals (i. e. the measured quantities), it is really an essential mathematical tool. Biologists and engineers even use the word "signal" in the sense of "function".

Periodic functions

A function f with real or complex values defined on the whole $\mathbb{R}$ is called a *periodic function* with period $T > 0$ iff for every $x \in \mathbb{R}$, it holds that $f(x+T) = f(x)$.

It is apparent that sums and scalar products of periodic functions with the same period are again periodic functions with the same period.

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 $x_0 \in \mathbb{R}$.

The last proposition can be proved easily:

Let us choose two such marginal points x_0 and y_0 for the integration Substituting $t = x + kT$ for a suitable k, we transform $\int_{y_0}^{y_0+T} f(x)\,dx$ to the case when $y_0 \in [x_0, x_0+T]$. Now we can split the interval of integration into three parts, thereby finishing the proof.

The orthogonality of the Fourier system of functions can be calculated by a nice trip to the world of complex numbers, which we can utilize later:

Let us remind that $e^{ix} = \cos(x) + i\sin(x)$. Straight differentiation of the product of real-valued functions, we can verify that real-valued functions $z(x)$ and $\varphi(x)$ of a real variable x satisfy

$$\big(z(x)\,e^{i\varphi(x)}\big)' = z'(x)\,e^{i\varphi(x)} + i\,z(x)\,\varphi'(x)\,e^{i\varphi(x)}\,.$$

$$= -\frac{1}{\pi}\int_0^{\pi} f(y)\left[\cos([2k]y)\cos([2k]\pi) - \sin([2k]y)\sin([2k]\pi)\right]\,dy$$

$$+\frac{1}{\pi}\int_0^{\pi} f(x)\cos([2k]x)\;dx$$

$$= -\frac{1}{\pi}\int_0^{\pi} f(y)\cos([2k]y)\;dy + \frac{1}{\pi}\int_0^{\pi} f(x)\cos([2k]x)\;dx = 0,$$

$$b_{2k} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin([2k]x)\;dx$$

$$= \frac{1}{\pi}\int_{-\pi}^{0} f(x)\sin([2k]x)\;dx + \frac{1}{\pi}\int_0^{\pi} f(x)\sin([2k]x)\;dx = |\,x = y+\pi\,|$$

$$= \frac{1}{\pi}\int_{-2\pi}^{-\pi} f(y+\pi)\sin([2k][y+\pi])\;dy + \frac{1}{\pi}\int_0^{\pi} f(x)\sin([2k]x)\;dx$$

$$= -\frac{1}{\pi}\int_0^{\pi} f(y)\sin([2k][y+\pi])\;dy + \frac{1}{\pi}\int_0^{\pi} f(x)\sin([2k]x)\;dx$$

$$= -\frac{1}{\pi}\int_0^{\pi} f(y)\left[\sin([2k]y)\cos([2k]\pi) + \sin([2k]\pi)\cos([2k]y)\right]\,dy$$

$$+\frac{1}{\pi}\int_0^{\pi} f(x)\sin([2k]x)\;dx$$

$$= -\frac{1}{\pi}\int_0^{\pi} f(y)\sin([2k]y)\;dy + \frac{1}{\pi}\int_0^{\pi} f(x)\sin([2k]x)\;dx = 0.$$

□

7.9. Decide the convergence and uniform convergence of the Fourier series of the function $g(x) = e^{-x}$ for $x \in [-1, 1)$.

Solution. We need not calculate the corresponding Fourier series if we only want to decide the convergence. Let us define a function s on $\mathbb{R}$ with period $T = 2$ as follows:

$$s(x) := g(x) = e^{-x},\ x \in (-1,1),\ s(1) := \frac{g(-1) + \lim\limits_{x\to 1-} g(x)}{2} = \frac{e+e^{-1}}{2}.$$

We know that this function is the sum of the Fourier series in question. In other words, the Fourier series converges to a periodic function s. 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 $[c, d] \cap \{2k+1;\ k \in \mathbb{Z}\} \neq \emptyset$ because a uniform limit of continuous functions is always a continuous function. Thus, the series converges to the function g on $(-1, 1)$, yet this convergence is uniform only on the subintervals $[c, d]$ which satisfy the restriction $-1 < c < d < 1$. □

7.10. Determine the cosine Fourier series for the periodic extension of the function

$$g(x) = 1,\ x \in [0, 1), \quad g(x) = 0,\ x \in [1, 4).$$

Further, determine 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. It is the case of the Fourier series of an even function. Therefore, the first thing we must do is to define the function g on the interval $(-4, 0)$ so that it is even, which means to set

$$g(x) := 1 \text{ for } x \in (-1, 0), \quad g(x) := 0 \text{ for } x \in (-4, -1].$$

The antiderivative of a complex function $f(x)$ of a real variable x can, of course, be obtained in terms of antiderivatives of the real and imaginary components of the function f.

Thus we can easily calculate the integral (supposing $m \neq n$)

$$\int_{-\pi}^{\pi} e^{imx}\,e^{-inx}\,dx = \int_{-\pi}^{\pi} e^{i(m-n)x}\,dx = \tfrac{1}{i(m-n)}[e^{i(m-n)x}]_{-\pi}^{\pi},$$

which always equals zero because it does not matter whether we run the multiples of π clockwise or counterclockwise.

The integral we have just determined expresses the scalar product $\langle e^{imx}, e^{inx}\rangle$. Thus we can see that all of our functions e^{inx} (with complex values) are, indeed, perpendicular to each other.

This scalar product can be rewritten as follows:

$$\begin{aligned}\langle e^{imx}, e^{inx}\rangle &= \langle \cos(mx) + i\sin(mx), \cos(nx) + i\sin(nx)\rangle \\ &= \big(\langle\cos(mx), \cos(nx)\rangle + \langle\sin(mx), \sin(nx)\rangle\big) \\ &\quad + i\big(\langle\sin(mx), \cos(nx)\rangle - \langle\cos(nx), \sin(mx)\rangle\big).\end{aligned}$$

We can notice that in the imaginary part of this expressions, there are odd functions integrated over the interval $[-\pi, \pi]$, which surely results in zero.

The functions $\sin(x)$ and $\cos(x)$ differ only by the phase shift, i. e. $\cos(mx - \pi/2) = \sin(mx)$. Thus both the summands in the real part of our expression are equal. Therefore, both of them must give zero. Thus we have verified the orthogonality of our system of functions.

At the same time, we can see that for $m = n$, the result is the real number $\int_{-\pi}^{\pi} dx = 2\pi$, and the sizes of both $\sin(nx)$ and $\cos(nx)$ must equal. Thus for positive numbers n, we necessarily get the sizes

$$\|\cos(nx)\|^2 = \pi, \quad \|\sin(nx)\|^2 = \pi.$$

For $n = 0$ only, we get $\|1\|^2 = 2\pi$.

Fourier series

A series of functions

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\big(a_n\cos(nx) + b_n\sin(nx)\big)$$

from the theorem 7.5, 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*.

In practice, we want to work with Fourier series with an arbitrary period T of the functions, not only the value 2π. Then it suffices to move to functions $\cos(\frac{2\pi}{T}nx)$, $\sin(\frac{2\pi}{T}nx)$. By mere substitution $t = \omega x$, where $\omega = \frac{2\pi}{T}$, we can verify the orthogonality of our new system of functions and recalculate the coefficients in the Fourier series $F(x)$ of a function g on the interval $[x_0, x_0 + T]$:

Now we can consider its periodic extension onto the whole $\mathbb{R}$ with period $T = 8$ and $\omega = \pi/4$.

We always must have $b_n = 0$ for all $n \in \mathbb{N}$ in a cosine Fourier series. We can also easily determine the Fourier coefficients

$$a_0 = \tfrac{2}{T} \int\limits_{x_0}^{x_0+T} g(x)\,dx = \tfrac{1}{2} \int\limits_{0}^{1} 1\,dx = \tfrac{1}{2},$$

$$a_n = \tfrac{2}{T} \int\limits_{x_0}^{x_0+T} g(x)\cos(n\omega x)\;dx = \tfrac{1}{2} \int\limits_{0}^{1} \cos\tfrac{n\pi x}{4}\,dx = \tfrac{2}{n\pi}\sin\tfrac{n\pi}{4},\ n \in \mathbb{N},$$

where we used the formula

$$\int\limits_{-a}^{a} f(x)\,dx = 2\int\limits_{0}^{a} f(x)\,dx, \tag{7.8}$$

which is valid for every even function f integrable on the interval $[0, a]$.

It is not a good idea to replace the expression $\sin(n\pi/4)$ in a similar way as in the previous exercises. We would have to divide the natural numbers n into 8 groups with respect to their remainder modulo 8. However, this would not yield much of a transparent expression. Thus we will do with the following form of the cosine Fourier series:

$$\tfrac{1}{4} + \sum_{n=1}^{\infty} \left[\tfrac{2}{n\pi}\sin\tfrac{n\pi}{4}\cos\tfrac{n\pi x}{4}\right].$$

The sine Fourier transform of the function can be determined analogously from the odd extension of the given segment. We again have $T = 8$ and $\omega = \pi/4$ for the function f. However, this time it is the coefficients a_n, $n \in \mathbb{N}\cup\{0\}$, which are zero. To find the remaining coefficient, we use integration by parts and (‖7.8‖) (the product of two odd functions is an even function), obtaining

$$\begin{aligned}
b_n &= \tfrac{2}{T} \int\limits_{x_0}^{x_0+T} f(x)\sin(n\omega x)\;dx \\
&= \tfrac{1}{2}\left[\int\limits_{0}^{2}(x-1)\sin\tfrac{n\pi x}{4}\,dx - \int\limits_{2}^{4}(x-3)\sin\tfrac{n\pi x}{4}\,dx\right] \\
&= \left[-(x-1)\tfrac{2}{n\pi}\cos\tfrac{n\pi x}{4}\right]_0^2 + \left[\tfrac{8}{n^2\pi^2}\sin\tfrac{n\pi x}{4}\right]_0^2 - \left[-(x-3)\tfrac{2}{n\pi}\cos\tfrac{n\pi x}{4}\right]_2^4 \\
&\quad - \left[\tfrac{8}{n^2\pi^2}\sin\tfrac{n\pi x}{4}\right]_2^4 = \tfrac{2}{n\pi}\left[(-1)^n - 1\right] + \tfrac{16}{n^2\pi^2}\sin\tfrac{n\pi}{2}, \quad n \in \mathbb{N}.
\end{aligned}$$

Hence we can immediately see that for even n, we have $b_n = 0$. Thanks to that, the sine Fourier series can be refined to the form

$$\begin{aligned}
&\sum_{n=1}^{\infty}\left[\left(\tfrac{2}{n\pi}\left[(-1)^n - 1\right] + \tfrac{16}{n^2\pi^2}\sin\tfrac{n\pi}{2}\right)\sin\tfrac{n\pi x}{4}\right] \\
&= \sum_{n=1}^{\infty}\left[\left(\tfrac{-4}{[2n-1]\pi} + \tfrac{(-1)^{n-1}16}{[2n-1]^2\pi^2}\right)\sin\tfrac{[2n-1]\pi x}{4}\right].
\end{aligned}$$

□

7.11. Express the function $g(x) = \cos x$, $x \in (0, \pi)$, as the sum of a cosine Fourier series and a sine Fourier series.

Solution. Of course, we have

$$\cos x = \cos x, \quad x \in (-\pi, \pi),$$

considering the left-hand cosine to be the even extension of the function g and the right-hand cosine to be the uniquely given cosine Fourier series.

Then, the sine series must have $a_n = 0$, $n \in \mathbb{N}\cup\{0\}$, and we can easily calculate that

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\bigl(a_n\cos(n\omega x) + b_n\sin(n\omega x)\bigr),$$

which have values

$$a_n = \frac{2}{T}\int_{x_0}^{x_0+T} g(x)\cos(n\omega x)\,dx,$$

$$b_n = \frac{2}{T}\int_{x_0}^{x_0+T} g(x)\sin(n\omega x)\,dx.$$

7.7. Exponential formula. A while ago, we used the basic formula for parametrization of the unit circle in the complex plane by the trigonometric functions when we verified the orthogonality of the functions $\cos(nx)$, $\sin(nx)$ If we consider $\omega = 2\pi/T$ to be the speed of running around the circle, where T is the time of one lap, we get the same parametrization in the form:

$$\mathrm{e}^{i\omega t} = \cos\omega t + i\sin\omega t.$$

For a (real or complex) function $f(t)$ and all integers n, we can define, in this context, its *complex Fourier coefficients* as the complex numbers

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-i\omega nt}\;dt.$$

Straight from the definition, 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 become clear. For natural numbers n, we get

$$c_n = \tfrac{1}{2}(a_n - ib_n), \quad c_{-n} = \tfrac{1}{2}(a_n + ib_n),$$

and if the function f takes on real values only, c_n and c_{-n} are, of course, complex conjugates of each other.

Thus we have expressed the Fourier series $F(t)$ for a function $f(t)$ in the form

$$F(t) = \sum_{n=-\infty}^{\infty} c_n\,\mathrm{e}^{i\omega nt}.$$

Both in the case of functions with real and complex values, the corresponding Fourier series can be written in this form. However, the coefficients will be complex in general in both cases.

We will return to this expression several times; for instance, when we will discuss the extraordinarily useful Fourier transformation.

We can notice that having fixed T, the expression $\omega = 2\pi/T$ describes just 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.

In subsequent parts of this chapter, we will show that Fourier series work with a complete orthogonal system on $\mathcal{S}^0$. However, we will have to prepare ourselves for that thoroughly. For that reason, we formulate some useful results right now and add some practically oriented notes. We will get back to the proofs later.

7.8. Theorem. *Let us consider a finite interval* $[a, b]$ *with length* $T = b-a$. *Further, let* f *be a function with real or complex values in* $\mathcal{S}^1[a, b]$ *(i. e. a piecewise continuous function with a piecewise continuous first derivative), periodically extended on the whole* $\mathbb{R}$. *Then:*

$$\begin{aligned} b_1 &= \tfrac{2}{\pi}\int_0^\pi \cos x \sin x\,dx = \tfrac{1}{\pi}\int_0^\pi \sin(2x)\;dx = 0,\\ b_n &= \tfrac{2}{\pi}\int_0^\pi \cos x \sin(nx)\;dx\\ &= \tfrac{1}{\pi}\int_0^\pi \sin([n+1]x) + \sin([n-1]x)\;dx\\ &= -\tfrac{1}{\pi}\left[\tfrac{\cos([n+1]x)}{n+1} + \tfrac{\cos([n-1]x)}{n-1}\right]_0^\pi = \tfrac{2n[(-1)^n+1]}{(n^2-1)\pi}, \qquad n \in \mathbb{N}\setminus\{1\}. \end{aligned}$$

Considering that

$$b_n = 0 \text{ for odd } n \in \mathbb{N} \quad\text{and}\quad b_n = \tfrac{4n}{(n^2-1)\pi} \text{ for even } n,$$

we get

$$\cos x = \sum_{n=1}^{\infty}\left[\tfrac{8n}{(4n^2-1)\pi}\sin(2nx)\right], \qquad x \in (0,\pi).$$

□

7.12. Write the Fourier series of the π-periodic function which equals cosine on the interval $(-\pi/2, \pi/2)$. Further, write the cosine Fourier series of the 2π-periodic function $y = |\cos x|$.

Solution. It is not hard to realize that we are looking for one Fourier series only (the second part of the problem is only reformulation of the first one). Therefore, let us construct the Fourier series for the function $g(x) = \cos x$, $x \in [-\pi/2, \pi/2]$. Since g is even, we have $b_n = 0$, $n \in \mathbb{N}$. At the same time, we have

$$\begin{aligned} a_0 &= \tfrac{2}{\pi}\int_{-\pi/2}^{\pi/2} \cos x\,dx = \tfrac{4}{\pi},\\ a_n &= \tfrac{2}{\pi}\int_{-\pi/2}^{\pi/2} \cos x \cos(2nx)\;dx\\ &= \tfrac{2}{\pi}\int_{-\pi/2}^{\pi/2} \tfrac{1}{2}\left[\cos([2n+1]x) + \cos([2n-1]x)\right]\,dx\\ &= \tfrac{1}{\pi}\left[\tfrac{\sin([2n+1]x)}{2n+1} + \tfrac{\sin([2n-1]x)}{2n-1}\right]_{-\pi/2}^{\pi/2} = \tfrac{2}{\pi}\left[\tfrac{(-1)^n}{2n+1} + \tfrac{(-1)^{n+1}}{2n-1}\right]\\ &= \tfrac{4}{\pi}\,\tfrac{(-1)^{n+1}}{4n^2-1} \end{aligned}$$

for every $n \in \mathbb{N}$. Let us notice that the calculation of a_0 could have been included in the calculation of a general coefficient a_n. The wanted Fourier series is

$$\tfrac{2}{\pi} + \tfrac{4}{\pi}\sum_{n=1}^{\infty}\left[\tfrac{(-1)^{n+1}}{4n^2-1}\cos(2nx)\right].$$

□

7.13. 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. All the way, we will use the formulae

(7.9)
$$\int \mathrm{e}^x\cos(\alpha x)\;dx = \frac{\mathrm{e}^x\left[\alpha\sin(\alpha x) + \cos(\alpha x)\right]}{1+\alpha^2} + C, \qquad \alpha \in \mathbb{R},$$

(7.10)
$$\int \mathrm{e}^x\sin(\beta x)\;dx = \frac{\mathrm{e}^x\left[\sin(\beta x) - \beta\cos(\beta x)\right]}{1+\beta^2} + C, \qquad \beta \in \mathbb{R},$$

which can be obtained by double integration by parts.

(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 piecewise continuous derivative, then the piecewise convergence of its Fourier series is uniform.

(3) The L_2–distance $\|s_N - f\|_2$ of the partial sums s_N of the Fourier series from the function f on $\mathcal{S}^1[a,b]$ always converges to zero for $N \to \infty$.

7.9. Extension of periodic functions. The convergent Fourier series converges, of course, outside the original interval $[-T/2, T/2]$ as well and it is a periodic function on the whole $\mathbb{R}$.

As an example, let us consider the Fourier series for the periodic function given by Heaviside's function $g(x)$ restricted to one period. I. e., our function g will be equal to -1 on the interval $[-\pi, 0]$ and to 1 on the interval $(0,\pi)$. We need not care about the values at zero and at the marginal points of the interval because these do not have any effect on the coefficients of the Fourier series. Its periodic extension onto the whole $\mathbb{R}$ is usually called an "angular wave function".

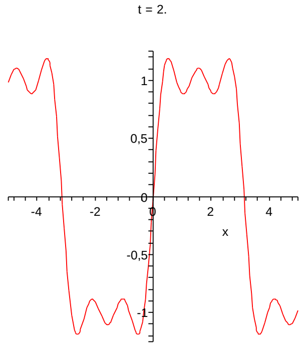
Since it is an odd function, the coefficients at the functions $\cos(nx)$ must be all zero. For the coefficients at the functions $\sin(nx)$, we get

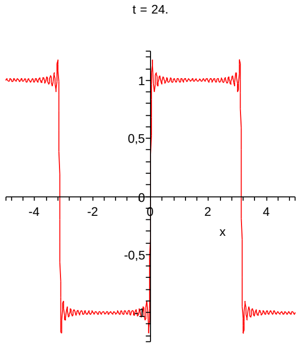
$$\begin{aligned} b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} g(x)\sin(nx)\,dx = \frac{2}{\pi}\int_0^{\pi}\sin(nx)\,dx\\ &= \frac{2}{n\pi}(1-(-1)^n). \end{aligned}$$

Thus the Fourier series has the form

$$g(x) = \frac{4}{\pi}\left(\sin(x) + \frac{1}{3}\sin(3x) + \frac{1}{5}\sin(5x) + \dots\right).$$

The partial sums of its first five and fifty terms, respectively, are shown in the following pictures.





If the interval $[-T/2, T/2]$ is chosen for the prime period of such an angular wave function, i. e. we want to work with the periodic extension of Heaviside's function with period T, we can easily recalculate that the resulting Fourier series has the form

$$g(x) = \frac{4}{\pi}\left(\sin(\omega x) + \frac{1}{3}\sin(3\omega x) + \frac{1}{5}\sin(5\omega x) + \dots\right),$$

where the number $\omega = \frac{2\pi}{T}$ is called the "phase frequency" of the wave. It expresses the ratio of the actual prime period to the unit frequency, i. e. the length 2π of the unit circle.

Thanks to them, we can get

(a)

$$\begin{aligned} a_0 &= 2\int\limits_0^1 \mathrm{e}^x\,dx = 2(\mathrm{e}-1),\\ a_n &= 2\int\limits_0^1 \mathrm{e}^x \cos(2n\pi x)\;dx = 2\left[\tfrac{\mathrm{e}^x[2n\pi\,\sin(2n\pi x)+\cos(2n\pi x)]}{1+4n^2\pi^2}\right]_0^1\\ &= \tfrac{2(\mathrm{e}-1)}{1+4n^2\pi^2},\quad n\in\mathbb{N},\\ b_n &= 2\int\limits_0^1 \mathrm{e}^x \sin(2n\pi x)\;dx = 2\left[\tfrac{\mathrm{e}^x[\sin(2n\pi x)-2n\pi\,\cos(2n\pi x)]}{1+4n^2\pi^2}\right]_0^1\\ &= \tfrac{4n\pi(1-\mathrm{e})}{1+4n^2\pi^2},\quad n\in\mathbb{N}; \end{aligned}$$

(b)

$$\begin{aligned} a_0 &= 2\int\limits_0^1 \mathrm{e}^x\,dx = 2(\mathrm{e}-1),\\ a_n &= 2\int\limits_0^1 \mathrm{e}^x \cos(n\pi x)\;dx = 2\left[\tfrac{\mathrm{e}^x[n\pi\,\sin(n\pi x)+\cos(n\pi x)]}{1+n^2\pi^2}\right]_0^1\\ &= \tfrac{2\left[(-1)^n\mathrm{e}-1\right]}{1+n^2\pi^2},\quad n\in\mathbb{N}; \end{aligned}$$

(c)

$$\begin{aligned} b_n &= 2\int\limits_0^1 \mathrm{e}^x \sin(n\pi x)\;dx = 2\left[\tfrac{\mathrm{e}^x[\sin(n\pi x)-n\pi\,\cos(n\pi x)]}{1+n^2\pi^2}\right]_0^1\\ &= \tfrac{2n\pi\left[1+(-1)^{n+1}\mathrm{e}\right]}{1+n^2\pi^2},\quad n\in\mathbb{N}. \end{aligned}$$

Straight substitution then yields the corresponding Fourier series

(a)

$$\mathrm{e}-1+2\,(\mathrm{e}-1)\sum_{n=1}^{\infty}\tfrac{\cos(2n\pi x)}{1+4n^2\pi^2}+4\pi\,(1-\mathrm{e})\sum_{n=1}^{\infty}\tfrac{n\,\sin(2n\pi x)}{1+4n^2\pi^2};$$

(b)

$$\mathrm{e}-1+2\sum_{n=1}^{\infty}\tfrac{\left[(-1)^n\mathrm{e}-1\right]\cos(n\pi x)}{1+n^2\pi^2};$$

(c)

$$2\pi\sum_{n=1}^{\infty}\tfrac{n\left[1+(-1)^{n+1}\mathrm{e}\right]\sin(n\pi x)}{1+n^2\pi^2}.$$

□

7.14. 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 up the series

$$\sum_{n=1}^{\infty}\tfrac{(-1)^{n+1}}{n^2},\qquad \sum_{n=1}^{\infty}\tfrac{1}{n^2}.$$

Solution. Once again, we could take advantage of the function g being even and calculate the non-zero coefficients a_n by integration by parts. However, in the theoretical part, the Fourier series for the function $f(x) = x^2$ on the interval $[-1, 1]$ is derived. This actually proves the identity

$$f(x) = \tfrac{1}{3} + \tfrac{4}{\pi^2}\sum_{n=1}^{\infty}\tfrac{(-1)^n\cos(n\pi x)}{n^2},\qquad x\in(-1,1).$$

Hence it follows (taking into account that $g(-\pi) = g(\pi)$) that

$$\begin{aligned} g(x) &= \pi^2 - \left(\tfrac{1}{3} + \tfrac{4}{\pi^2}\sum_{n=1}^{\infty}\tfrac{(-1)^n\cos\frac{n\pi x}{\pi}}{n^2}\right)\pi^2\\ &= \tfrac{2}{3}\pi^2 + 4\sum_{n=1}^{\infty}\tfrac{(-1)^{n+1}\cos(nx)}{n^2},\qquad x\in[-\pi,\pi]. \end{aligned}$$

It sufficed to add π^2 and multiply the original series by -1. Further, we have to realize that only nx, and not $n\pi x$, will be the arguments of the cosines. Thus the period is π times as great ($2/T$ and the integration bounds in the formula for a_n are changed), and integrating the cosines now does

Let us notice that as the number of terms of the series increases, the approximation gets much better except in a (still shrinking) neighborhood of the discontinuity point. There, the maximum of the deviance remains roughly the same. This is a general property of Fourier series and it is called *Gibbs phenomenon*.

Let us also notice that at the point of discontinuity, the value of the approximating function right half between the one-sided limits for Heaviside's function, just like 7.8(1) claims.

Of course, we cannot expect that the convergence of Fourier series for functions g with discontinuity points be uniform (then, the function g would have to be continuous itself, being a uniform limit of continuous functions).

7.10. Utilizing symmetry of functions. Let us think about the problem how we could approximate the function $g(x) = x^2$ by a Fourier series on the interval $[0, 1]$ as best as possible. If we just periodically extended this function from the given interval $[0, 1]$, it would not be continuous, and so the convergence at integers would be as queer as in the case of an angular wave function. However, we can easily work with the Fourier series on the base interval $[-1, 1]$. It is an even function on this domain, and so only the coefficients a_n can be non-zero.

For $n > 0$, double application of integration by parts yields:

$$\begin{aligned} a_n &= \frac{2}{2}\int_{-1}^{1} x^2\cos\!\left(\tfrac{2\pi nx}{2}\right)dx = 2\int_0^1 x^2\cos(\pi nx)dx\\ &= \frac{4}{\pi^2 n^2}(-1)^n. \end{aligned}$$

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

$$f(x) = \frac{1}{3} + \frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos(\pi nx).$$

By Weierstrass criterion, it is apparent that this series converges uniformly. Therefore, $f(x)$ will be continuous. Thanks to the theorem 7.8, we know that actually $f(x) = x^2$ on the whole interval $[-1, 1]$ since we are approximating a continuous function on the whole $\mathbb{R}$, and the convergence must be uniform. Thus our series approximates the function x^2 on the interval $[0, 1]$ far better than we could do it with a periodic extension of the function from this interval only.

Let us proceed with our illustrations. Thanks to the uniform convergence, we can invoke the rule for differentiating and integrating series term by term and calculate the Fourier series for the functions x and x^3. The differentiation will be the simpler one:

$$\frac{1}{2}(x^2)' = x = \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin(\pi nx).$$

Apparently, this series cannot converge uniformly since the periodic extension of the function x is not a continuous function. However, we can easily derive that it will converge pointwise (see our reasonings about alternating series in **??**), thus we really obtained the equality. (see the theorem **??**).

not give π in the denominators (the change of the upper bound takes effect when calculating a_0). Therefore, we had to multiply the original series by π^2. Readers who are unable to perform the corresponding calculations in mind and to immediately realize the differences, are advised to calculate the Fourier series of the function g directly.

Substituting $x = 0$ and $x = \pi$ then gives

$$\pi^2 = \tfrac{2}{3}\pi^2 + 4\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n^2}, \quad \text{i. e.} \quad \sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}}{n^2} = \tfrac{\pi^2}{12},$$

and

$$0 = \tfrac{2}{3}\pi^2 + 4\sum_{n=1}^{\infty} \tfrac{(-1)^{n+1}(-1)^n}{n^2}, \quad \text{i. e.} \quad \sum_{n=1}^{\infty} \tfrac{1}{n^2} = \tfrac{\pi^2}{6}.$$

In other words, we have found another way of expressing

$$\pi^2 = 12\left(1 - \tfrac{1}{2^2} + \tfrac{1}{3^2} - \tfrac{1}{4^2} + \cdots\right) = 6\left(1 + \tfrac{1}{2^2} + \tfrac{1}{3^2} + \tfrac{1}{4^2} + \cdots\right).$$

□

7.15. Using the Fourier series of the function $g(x) = \mathrm{e}^x, x \in [0, 2\pi)$, calculate $\sum_{n=1}^{\infty} \frac{1}{1+n^2}$.

Solution. We have (see also (||7.9||), (||7.10||))

$$\begin{aligned} a_0 &= \tfrac{1}{\pi}\int_0^{2\pi} \mathrm{e}^x \, dx = \tfrac{1}{\pi}\left(\mathrm{e}^{2\pi} - 1\right), \\ a_n &= \tfrac{1}{\pi}\int_0^{2\pi} \mathrm{e}^x \cos(nx) \, dx = \tfrac{1}{\pi}\left[\tfrac{\mathrm{e}^x[\cos(nx)+n\sin(nx)]}{1+n^2}\right]_0^{2\pi} \\ &= \tfrac{\mathrm{e}^{2\pi}-1}{(1+n^2)\pi}, \quad n \in \mathbb{N}, \\ b_n &= \tfrac{1}{\pi}\int_0^{2\pi} \mathrm{e}^x \sin(nx) \, dx \\ &= \tfrac{1}{\pi}\left[\tfrac{\mathrm{e}^x[\sin(nx)-n\cos(nx)]}{1+n^2}\right]_0^{2\pi} = -\tfrac{n(\mathrm{e}^{2\pi}-1)}{(1+n^2)\pi}, \quad n \in \mathbb{N}. \end{aligned}$$

Therefore,

$$\mathrm{e}^x = \tfrac{\mathrm{e}^{2\pi}-1}{\pi}\left(\tfrac{1}{2} + \sum_{n=1}^{\infty} \tfrac{\cos(nx)-n\sin(nx)}{1+n^2}\right), \quad x \in (0, 2\pi).$$

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 apparently not continuous at this point, so we get

$$\tfrac{\mathrm{e}^0+\mathrm{e}^{2\pi}}{2} = \tfrac{g(0)+\lim\limits_{x\to 2\pi-} g(x)}{2} = \tfrac{\mathrm{e}^{2\pi}-1}{\pi}\left(\tfrac{1}{2} + \sum_{n=1}^{\infty} \tfrac{\cos 0 - n\sin 0}{1+n^2}\right),$$

hence it follows that

$$\tfrac{\mathrm{e}^{2\pi}+1}{2} \cdot \tfrac{\pi}{\mathrm{e}^{2\pi}-1} = \tfrac{1}{2} + \sum_{n=1}^{\infty} \tfrac{1}{1+n^2}$$

which can be refined to

$$\sum_{n=1}^{\infty} \tfrac{1}{1+n^2} = \tfrac{(\pi-1)\,\mathrm{e}^{2\pi}+\pi+1}{2(\mathrm{e}^{2\pi}-1)}.$$

□

7.16. Calculate the series

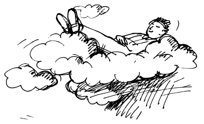
$$\sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^2}.$$

Solution. To determine the value of this series, one can successfully apply many known Fourier series of various functions. Let us remind, for instance, the Fourier series

Similarly, we can integrate term by term, leading to

$$\frac{1}{3}x^3 = \frac{2}{3}x + \frac{4}{\pi^3}\sum_{n=1}^{\infty} \frac{(-1)^n}{n^3}\sin(\pi n x),$$

and the resulting Fourier series is obtained by substituting for x from the previous equality.



7.11. General Fourier series and wavelets. In the case of a general orthogonal system of functions f_n and the series made from it, we often talk about *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, pictures, and so on. The nature of the period trigonometric functions used in the standard Fourier series and their simple scaling by increasing the frequency limit their usability. In many application fields, there arose a natural requirement of more convenient orthogonal systems of functions which will reflect the supposed nature of the data and which could be processed more efficiently.

Requirements for fast numerical processing usually include quick scalability and the possibility of easy movements by constant values. We can hope in such a system if, for instance, we choose a suitable continuous function ψ with a compact support from which we create countably many functions ψ_{jk}, $j, k \in \mathbb{Z}$, by translations and dilations:

$$\psi_{jk}(x) = 2^{j/2}\,\psi(2^j x - k).$$

If at the same time, the following two conditions are satisfied:

- the form of the *mother function* ψ captures the possible behavior of the data in a good way,
- its descendants ψ_{jk} form a complete orthogonal system,

then, possibly, only a few of the functions will do to approximate the processed signal in question. We talk about the so-called *wavelets*.

We have no space for details here, it is an extraordinarily vital field of research as well as the base of commercial applications. Interested readers are referenced to special literature.

Let us remark that actually, only discrete versions of our objects are used, i. e. the values of all the functions ψ_{jk} are only enlisted in a discrete (very large) set of points and are also orthogonal in this sense. The standards JPEG2000 are a good example of this, they use this technique and are a tool for professional compression of visual data in film industry, or the format DjVu for compressed publications.

One of the first wavelets was created by Ingrid Daubechies. In the picture below, there are the so-called Daubechies mother wavelet $D4(x)$ and its daughter $D4(2^{-3}x - 1)$.

$$\frac{\pi}{2}-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\cos([2n-1]x)}{(2n-1)^2},$$

which was calculated for the function $g(x)=|\,x\,|, x\in[-\pi,\pi)$. Since this function is continuous on $[-\pi,\pi)$ and $|-\pi|=|\pi|$, we even know that

$$|\,x\,|=\frac{\pi}{2}-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\cos([2n-1]x)}{(2n-1)^2},\qquad x\in[-\pi,\pi].$$

Substituting $x=0$ gives

$$0=\frac{\pi}{2}-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{(2n-1)^2},\qquad \text{i. e.}\qquad \sum_{n=1}^{\infty}\frac{1}{(2n-1)^2}=\frac{\pi^2}{8}.$$

□

7.17. Sum up the series

$$\sum_{n=1}^{\infty}\frac{1}{n^4},\qquad \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^4}.$$

Solution. First, let us remind that the values of the series

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6},\qquad \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2}=\frac{\pi^2}{12}$$

have already been determined. In this exercise, we will hint the procedure by which the series

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k}},\qquad \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{2k}}$$

for a general $k\in\mathbb{N}$ can be calculated. We use the identities

(7.11) $$x=\pi-2\sum_{n=1}^{\infty}\frac{\sin(nx)}{n},\qquad x\in(0,2\pi),$$

(7.12)
$$x^2=\frac{4\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{\cos(nx)}{n^2}-4\pi\sum_{n=1}^{\infty}\frac{\sin(nx)}{n},\qquad x\in(0,2\pi),$$

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 (‖7.11‖), we have

$$\sum_{n=1}^{\infty}\frac{\sin(nx)}{n}=\frac{\pi-x}{2},\qquad x\in(0,2\pi).$$

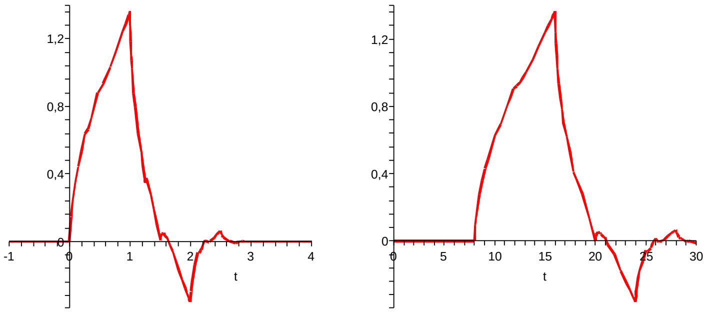
Substituting into (‖7.12‖) gives

$$\sum_{n=1}^{\infty}\frac{\cos(nx)}{n^2}=\frac{3x^2-6\pi x+2\pi^2}{12},\qquad x\in(0,2\pi).$$

Mere substitution then proves the validity of this formula at the marginal points $x=0$, $x=2\pi$ as well. The left-hand series is apparently bounded from above by $\sum_{n=1}^{\infty}\frac{1}{n^2}$, thus it converges absolutely and uniformly on $[0,2\pi]$. Therefore, it can be integrated term by term:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{\sin(nx)}{n^3}&\\ &=\sum_{n=1}^{\infty}\left[\frac{\sin(ny)}{n^3}\right]_0^x=\int_0^x\sum_{n=1}^{\infty}\frac{\cos(ny)}{n^2}\,dy\\ &=\int_0^x\frac{3y^2-6\pi y+2\pi^2}{12}\,dy=\frac{x^3-3\pi x^2+2\pi^2x}{12},\quad x\in[0,2\pi].\end{aligned}$$

Let us point out that, in fact, every Fourier series may be integrated term by term. Similarly, further integration gives



The function $F4$ is not described by any means from analysis. The function is given only by the values it takes on a finite (yet very large) set of input values. It is chosen so that it would have, in its various parts, all the necessary properties for graphics data — both slow and fast increase, sharp turns at both extrema, and so on. The complexity of the construction lies, of course, in the condition that the system obtained by the mentioned construction be orthogonal!

2. Metric spaces

In this part of the chapter, we will focus on the concepts of distance and convergence in a more abstract way. We will take advantage of this presently when we prove the already mentioned properties of Fourier series, and we will also return to these concepts in miscellaneous contexts. So we can consider the subsequent pages to be a very useful (and hopefully manageable) trip into the world of mathematics for the competent or courageous.

7.12. Metrics and norms. When we derived the technique of Fourier series, we freely talked about the distance on a space of functions. Now we will stop by this concept and explain it thoroughly.

The Euclidean distance in the vector spaces $\mathbb{R}^n$ satisfies the following three abstract requirements. (So does the L_1–distance $d(f,g)=\|f-g\|_1$ on the space of continuous and absolutely integrable functions.) For the oncoming paragraphs, let us try to keep these two examples in our minds.

Axioms of a metric and a norm

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

(7.2) $$d(x,y)\geq 0;\text{ and } d(x,y)=0 \text{ iff } x=y\ ,$$

(7.3) $$d(x,y)=d(y,x),$$

(7.4) $$d(x,z)\leq d(x,y)+d(y,z),$$

is called a *metric space*. The mapping d is the *metric* on X.

If X is a vector space over $\mathbb{R}$ and $\|\ \| : X\to\mathbb{R}$ is a function satisfying

(7.5) $$\|x\|\geq 0;\text{ and } \|x\|=0 \text{ iff } x=0\ ,$$

(7.6) $$\|\lambda x\|=|\lambda|\,\|x\|,\text{ for all scalars }\lambda\ ,$$

(7.7) $$\|x+y\|\leq\|x\|+\|y\|,$$

then the function $\|\ \|$ is called the *norm* on X, and the space X is then a *normed vector space*.

A norm always gives the metric $d(x,y)=\|x-y\|$.

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1-\cos(nx)}{n^4} &= \sum_{n=1}^{\infty}\left[-\frac{\cos(ny)}{n^4}\right]_0^x = \int_0^x\sum_{n=1}^{\infty}\frac{\sin(ny)}{n^3}\,dy\\ &= \int_0^x\frac{y^3-3\pi y^2+2\pi^2 y}{12}\,dy = \frac{x^4-4\pi x^3+4\pi^2 x^2}{48}, \qquad x\in[0,2\pi].\end{aligned}$$

Substituting $x=\pi$ leads to

$$\sum_{n=1}^{\infty}\frac{1+(-1)^{n+1}}{n^4} = \sum_{n=1}^{\infty}\frac{1-\cos(n\pi)}{n^4} = \frac{\pi^4}{48}.$$

Taking into account the fact the left-hand numerator is zero for even numbers n and equals 2 for odd numbers n, the obtained series can be written as

$$(7.13)\qquad \sum_{n=1}^{\infty}\frac{2}{(2n-1)^4} = \frac{\pi^4}{48}.$$

From the expression

$$\sum_{n=1}^{\infty}\frac{1}{n^4} = \sum_{n=1}^{\infty}\frac{1}{(2n)^4} + \sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} = \frac{1}{16}\sum_{n=1}^{\infty}\frac{1}{n^4} + \sum_{n=1}^{\infty}\frac{1}{(2n-1)^4},$$

it follows that

$$\sum_{n=1}^{\infty}\frac{1}{n^4} = \frac{16}{15}\sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} = \frac{16}{15}\cdot\frac{1}{2}\cdot\frac{\pi^4}{48} = \frac{\pi^4}{90},$$

thereby having summed up the first series. As for the second one, we have

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^4} &= \sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} - \sum_{n=1}^{\infty}\frac{1}{(2n)^4} = \sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} - \frac{1}{16}\sum_{n=1}^{\infty}\frac{1}{n^4}\\ &= \frac{1}{2}\cdot\frac{\pi^4}{48} - \frac{1}{16}\cdot\frac{\pi^4}{90} = \frac{7\pi^4}{720}.\end{aligned}$$

As we have said, one can proceed similarly when summing up the series

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k}},\quad \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{2k}}$$

for other $k\in\mathbb{N}$. Therefore, it is natural to ask for the value of the series $\sum_{n=1}^{\infty}\frac{1}{n^3}$. This problem has been tackled by mathematicians for centuries without success. The reader may justifiably be surprised by this since the mentioned procedure should be applicable to all the odd powers as well.

We can, for instance, start with the identity

$$\sum_{n=1}^{\infty}\frac{\cos(nx)}{n} = -\ln\left(2\sin\tfrac{x}{2}\right), \quad x\in(0,2\pi),$$

which, by the way, can be proved by expanding the right-hand function into a Fourier series again. If we, similarly to above, integrated the left-hand series term by term twice and substituted $x\to 0+$ in the limit, we would get just the series $\sum_{n=1}^{\infty}\frac{1}{n^3}$. Thus, it should suffice to integrate the right-hand function twice and calculate one limit. However, the integration of the right-hand side leads to a non-elementary integral, i. e., the antiderivative cannot be expressed in terms of elementary function we usually work with. [1]

□

7.18. Using Parseval's identity for Fourier's orthogonal system, verify that

$$\sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} = \frac{\pi^4}{96}.$$

At the beginning of the previous part of this chapter, we actually defined the distance of functions using the so-called L_1–norm. In Euclidean vector spaces, it was the norm $\|x\|$, which is induced by the bilinear scalar product by the relation $\|x\|^2=\langle x,x\rangle$. Similarly, we worked with the norm on unitary spaces. We obtained the L_2–norm on continuous functions in the same way.

Of course, metrics given by a norm have very specific properties since their behavior on the whole space X can be derived from the properties in an arbitrarily small neighborhood of the zero element $x=0\in X$.

7.13. Convergence. The concepts of (close) neighborhoods of particular elements, convergence of sequences of elements and the corresponding "topological" concepts can be defined on completely abstract metric spaces in much the same way as in the case of the real and complex numbers and their sequences at the beginning of the fifth chapter, see 5.12–5.17.

We can almost copy these paragraphs; only the proofs of the theorem 5.17 will be much harder. We will start off with the concept of convergent sequences in a metric space X with metric d:

Cauchy sequences

Let us consider an arbitrary sequence of elements $x_0, x_1, \dots$ lying in X such that for any fixed positive real number ε, it holds for all but finitely many terms x_i of the sequence that for all but finitely many terms x_j,

$$d(x_i,x_j)<\varepsilon.$$

In other words, for any fixed $\varepsilon>0$, there is an index N such that the above inequality holds for all $i,j>N$; i. e. 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 value x in the following sense:

Convergent sequences

If a sequence of elements $x_0, x_1, \dots \in X$, a fixed element $x\in X$ and every positive real number ε are such that for all but finitely many i (depending on the choice of ε), it holds that

$$d(x_i,x)<\varepsilon,$$

we say that the sequence x_i, $i=0,1,\dots$, *converges* to the element x, which is called the *limit* of the sequence x_i, $i=0,1,\dots$ in the metric space X.

Thanks to the triangle inequality, we get that for each pair of terms x_i, x_j from a convergent sequence with sufficiently large indeces, it holds that (the denotation is taken from the definition above)

$$d(x_i,x_j)\le d(x_i,x)+d(x,x_j)<2\varepsilon.$$

Therefore, every convergent sequence is a Cauchy sequence. Metric spaces where the converse (i. e. *every Cauchy sequence is convergent*) is true as well are called *complete metric spaces*.

7.14. Topology, convergence, and continuity. Just as in the case of the real numbers, we can formulate the convergence in terms of "open neighborhoods".

[1] The function $\zeta(p)=\sum_{n=1}^{\infty}$ is called the Riemann zeta function.

Solution. We have already summed this series up (see (‖7.13‖)). Now, we will reveal that number series can be summed up even more easily thanks to Fourier series. However, this way is conditioned by knowledge of a good deal of Fourier series and can be a bit more complicated for the reader. (We recommend to compare the solutions of this exercise and the previous one.)

It is imperative to choose an appropriate Fourier series. For instance, let us take the Fourier series

$$\tfrac{\pi}{2} - \tfrac{4}{\pi} \sum_{n=1}^{\infty} \tfrac{\cos([2n-1]x)}{(2n-1)^2},$$

which we have obtained for the function $g(x) = |\,x\,|, x \in [-\pi, \pi)$ and which has already been used once to determine the value of a series. Parseval's identity

$$\tfrac{a_0^2}{2} + \sum_{n=1}^{\infty} a_n^2 + \sum_{n=1}^{\infty} b_n^2 = \tfrac{2}{T} \int\limits_{x_0}^{x_0+T} [g(x)]^2\, dx$$

says for it that

$$\tfrac{\pi^2}{2} + \tfrac{16}{\pi^2} \sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} = \tfrac{1}{\pi} \int\limits_{-\pi}^{\pi} |\,x\,|^2\, dx = \tfrac{2}{\pi} \int\limits_{0}^{\pi} x^2\, dx = \tfrac{2\pi^2}{3},$$

i. e.,

$$\sum_{n=1}^{\infty} \tfrac{1}{(2n-1)^4} = \left(\tfrac{2\pi^2}{3} - \tfrac{\pi^2}{2}\right) \tfrac{\pi^2}{16} = \tfrac{\pi^4}{96}.$$

□

Now, we will illustrate how Fourier series can be applied in the theory of differential equations. For the sake of simplicity, we consider only the non-homogeneous (compare to (‖7.2‖)) differential equation

$$y'' + a^2 y = f(x) \tag{7.14}$$

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 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., the identity

$$f(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \left[A_n \cos \frac{2\pi n x}{T} + B_n \sin \frac{2\pi n x}{T} \right], \quad x \in \mathbb{R}. \tag{7.15}$$

7.19. Prove that if the equation (‖7.14‖) 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 (‖7.14‖) has a unique periodic solution with period T if and only if

$$a \neq \frac{2\pi n}{T} \quad \text{for every } n \in \mathbb{N}. \tag{7.16}$$

Solution. Let a function $y = g(x)$, $x \in \mathbb{R}$, be a solution of the equation (‖7.14‖) and let it have 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'', ... share the same period, the function

$$g''(x) + a^2 g(x) = f(x)$$

is also period 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 $l \in \mathbb{N}$.

Open and closed sets

The *open ε–neighborhood* of an element x in a metric space X (or just ε–neighborhood in short) is the set

$$\mathcal{O}_\varepsilon(x) = \{y \in X;\ d(x, y) < \varepsilon\}.$$

A subset $U \subset X$ is *open* iff with every point x it contains, it contains some ε–neighborhood of x as well. A subset $W \subset X$ is *closed* iff its complement $X \setminus W$ is an open set.

Instead of an ε–neighborhood, we also talk about an (open) ε–ball centered at x. In the case of a normed space, we can do with ε–balls centered at zero: those added to the given element x give just its ε–neighborhood.

The *limit points of a subset $A \subset X$* are, again, defined as such elements $x \in X$ that there is a sequence of points from A converging to, but not containing x. We can easily see that a set is closed if and only if it contains all of its limit points:

Indeed, it follows straight from the definition that a set A is closed if and only if for every point $x \notin A$, there is some $\varepsilon > 0$ such that the whole ε–neighborhood $\mathcal{O}_\varepsilon(x)$ has an empty intersection with A. Thus if A were closed and x were a limit point of the set A not belonging to it, then in every such ε–neighborhood of such a point x, there are infinitely many points of the set A, which is a contradiction.

On the other hand, let us suppose that A contains all of its limit points and let us consider $x \in X \setminus A$. If in every ε–neighborhood of the point x, there were a point $x_\varepsilon \in A$, then the choices $\varepsilon = 1/n$ give us 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 in A which belong to A together with some neighborhood of theirs. Further, we define the closure $\bar{A}$ of a set A as the union of the original set A with the set of all A's limit points.

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 results in a closed set again.

It is the other case with open sets: any union of open sets is an open set, but in general, only 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, while the closure of A is the intersection of all closed sets which contain A.

The closed and open sets are the essential concepts of the mathematical discipline called *topology*. Without going into deeper connections, we have just made ourselves familiar with the *topology of the metric spaces*.

The concept of convergence can be reformulated now as follows: a sequence of elements x_i, $i = 0, 1, \dots$, in a metric space X converges to $x \in X$ iff 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:

A mapping $f : W \to Z$ is continuous iff the reverse image $f^{-1}(V)$ of every open set $V \subset Z$ is an open set in W. Of course, this means nothing else than the claiming that for every $z = f(x) \in Z$ and a positive real number ε, there is a positive real number δ

Now, suppose that the a function $y = g(x)$, $x \in \mathbb{R}$, is a periodic solution of the equation (‖7.14‖) with period T and that it is expressed by a Fourier series as follows:

$$(7.17)\quad g(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(\omega n x) + b_n \sin(\omega n x)\right], \quad x \in \mathbb{R},$$

where $\omega = 2\pi/T$. If g satisfies the equation (‖7.14‖), it must have a continuous second derivative on $\mathbb{R}$. Therefore,

$$g'(x) = \sum_{n=1}^{\infty} \left[\omega n b_n \cos(\omega n x) - \omega n a_n \sin(\omega n x)\right], \quad x \in \mathbb{R},$$

(7.18)

$$g''(x) = \sum_{n=1}^{\infty} \left[-\omega^2 n^2 a_n \cos(\omega n x) - \omega^2 n^2 b_n \sin(\omega n x)\right], \quad x \in \mathbb{R}.$$

Substituting (‖7.15‖), (‖7.17‖) and (‖7.18‖) into (‖7.14‖) yields

$$a^2 \tfrac{a_0}{2} + \sum_{n=1}^{\infty} \left[\left(-\omega^2 n^2 a_n + a^2 a_n\right) \cos(n\omega x) + \left(-\omega^2 n^2 b_n + a^2 b_n\right) \sin(n\omega x)\right]$$

$$= \tfrac{A_0}{2} + \sum_{n=1}^{\infty} \left[A_n \cos(n\omega x) + B_n \sin(n\omega x)\right].$$

Hence it follows that

$$(7.19)\qquad a^2 \frac{a_0}{2} = \frac{A_0}{2}, \quad \text{i. e.} \quad a_0 = \frac{A_0}{a^2},$$

and

(7.20)

$$\left(-\omega^2 n^2 + a^2\right) a_n = A_n, \quad \left(-\omega^2 n^2 + a^2\right) b_n = B_n, \quad n \in \mathbb{N}.$$

We can see that 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 (‖7.16‖) holds. In this case, the only solution of (‖7.14‖) with period T is determined by the only solution

$$(7.21)\qquad a_n = \frac{A_n}{-\omega^2 n^2 + a^2}, \quad b_n = \frac{B_n}{-\omega^2 n^2 + a^2}, \quad n \in \mathbb{N}$$

of the system (‖7.20‖). Let us emphasize that we have silently utilized the uniform convergence of the series in (‖7.18‖). This follows, besides others, from deeper results of the general theory of Fourier series to which we will not pay further attention. □

7.20. Using the solution of the previous problem, find all 2π-periodic solutions of the differential equation

$$y'' + 2y = \sum_{n=1}^{\infty} \tfrac{\sin(nx)}{n^2}, \quad x \in \mathbb{R}.$$

Solution. The equation is in the form of (‖7.14‖) for $a = \sqrt{2}$ and the continuously differentiable function

$$f(x) = \sum_{n=1}^{\infty} \tfrac{\sin(nx)}{n^2}, \quad x \in \mathbb{R}$$

with prime period $T = 2\pi$. According to the problem ‖7.19‖, the condition $\sqrt{2} \notin \mathbb{N}$ implies the there is exactly one 2π-periodic solution. If we look for it as the value of the series

$$\tfrac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx)\right], \quad x \in \mathbb{R},$$

such that for all elements $y \in W$ with distance $d_W(x, y) < \delta$, it also holds that $d_Z(z, f(y)) < \varepsilon$.

Again, similarly as in the case of the real-valued functions, a mapping f between metric spaces is continuous if and only if it respects convergence of sequences.

7.15. L_p–norms. Now we have the general tools with which we can have a look at examples of metric spaces created by finite-dimensional vectors or functions at our disposal. We will restrict ourselves to an extraordinarily useful class of norms.

We begin with the real or complex finite-dimensional vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$, and for a fixed real number $p \geq 1$ and any vector $z = (z_1, \dots, z_n)$, we define

$$\|z\|_p = \left(\sum_{i=1}^{n} |z_i|^p\right)^{1/p}.$$

We are going to 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 will use the so-called *Hölder's 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, it holds that*

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n} y_i^q\right)^{1/q},$$

where $1/q = 1 - 1/p$.

Proof. Let us 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 clearly holds. Therefore, let us suppose that $X \neq 0$ and $Y \neq 0$.

Hölder's inequality is a useful straight corollary of the convexity of the exponential function. Let us define the numbers v_k and w_k so that

$$x_k = X\, \mathrm{e}^{v_k/p}, \quad y_k = Y\, \mathrm{e}^{w_k/q}.$$

Since $1/p + 1/q = 1$, we can consider the affine combination of the values $\frac{1}{p} v_k + \frac{1}{q} w_k$ and thanks to the mentioned convexity, we obtain

$$\mathrm{e}^{v_k/p + w_k/q} \leq \tfrac{1}{p}\, \mathrm{e}^{v_k} + \tfrac{1}{q}\, \mathrm{e}^{w_k}.$$

Hence we can calculate straightaway that

$$\frac{1}{XY} x_k y_k \leq \frac{1}{p}\left(\frac{x_k}{X}\right)^p + \frac{1}{q}\left(\frac{y_k}{Y}\right)^q,$$

and summing over $k = 1, \dots, n$,

$$\frac{1}{XY} \sum_{i=1}^{n} x_i y_i \leq \frac{1}{pX^p} \sum_{i=1}^{n} x_i^p + \frac{1}{qY^p} \sum_{i=1}^{n} y_i^q.$$

However, the particular sums on the right-hand side give exactly X^p and Y^q, so the whole expression is equal to $1/p + 1/q = 1$. Multiplying this inequality by the number XY finishes the proof. □

Now we will really be able to prove that $\|\ \|_p$ is indeed a norm:

we further know that (see (‖7.19‖) and (‖7.21‖))

$$a_0 = a_n = 0, \quad b_n = \frac{1}{n^2(2-n^2)}, \quad n \in \mathbb{N}.$$

Thus, the given equation has the unique 2π-periodic solution

$$y = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2(2-n^2)}, \quad x \in \mathbb{R}.$$

□

C. Metric spaces

7.21. Show that the definition of a metric as a function d defined on $X \times X$ for a non-empty set X and satisfying

(7.22) $\qquad d(x, y) = 0$, if and only if $x = y$, $\quad x, y \in X$,

(7.23) $\qquad d(x, z) \le d(y, x) + d(y, z), \quad x, y, z \in X$,

is equivalent to the definition given in the theoretical part, in paragraph 7.12.

Solution. Ostensibly, this definition lays fewer requirements on the metric than the definition from the theoretical part. The definitions are equivalent iff the conditions (‖7.22‖), (‖7.23‖) imply

(7.24) $\qquad d(y, x) \ge 0, \quad x, y \in X$,

(7.25) $\qquad d(x, y) = d(y, x), \quad x, y \in X$.

However, if we set $x = z$ in (‖7.23‖), we get (‖7.24‖) from (‖7.22‖). Similarly, the choice $y = z$ in (‖7.23‖), using (‖7.22‖), implies that $d(x, y) \le d(y, x)$ for all points $x, y \in X$. Interchanging the variables x and y then gives $d(y, x) \le d(x, y)$, i. e. (‖7.25‖). Thus, we have proved that the definitions are equivalent.

Many more ways of defining a metric can be found in literature. Besides those, one can find many definitions which are a bit different and lead to objects other than metrics (the most important ones being pseudometrics, ultrametrics, and semimetrics). 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. □

7.22. Consider the power set (the set of all subsets) of a given finite set. Determine whether the mapping defined for all considered subsets X, Y by

(a) $d_1(X, Y) := |\,(X \cup Y) \smallsetminus (X \cap Y)\,|$;

(b) $d_2(X, Y) := \frac{|\,(X\cup Y)\smallsetminus(X\cap Y)\,|}{|\,X\cup Y\,|}, \quad X \cup Y \ne \emptyset, \quad d_2(\emptyset, \emptyset) := 0$

is a metric. (By $|\,X\,|$, we mean the number of elements of a set X.)

Solution. We will 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. The reader should immediately realize that both d_1 and d_2 satisfy them. Therefore, we analyze the triangle inequality only.

The case (a). For any sets X, Y, Z, we have

(7.26)
$$(X \cup Z) \smallsetminus (X \cap Z) \subseteq [(X \cup Y) \smallsetminus (X \cap Y)] \cup [(Y \cup Z) \smallsetminus (Y \cap Z)]$$

since if $x \in (X \cup Z) \smallsetminus (X \cap Z)$, then exactly one of the following occurs:

$$x \in X \text{ and } x \notin Z, \quad x \notin X \text{ and } x \in Z.$$

Minkowski inequality

For every $p > 1$ and all n–tuples of non-negative real numbers $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$, it holds that

$$\left(\sum_{i=1}^{n}(x_i + y_i)^p\right)^{1/p} \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} + \left(\sum_{i=1}^{n} y_i^p\right)^{1/p}.$$

To verify this practical inequality, we can use the following trick, invoking Hölder's inequality. We surely have (notice that $p > 1$)

$$\sum_{i=1}^{n} x_i(x_i + y_i)^{p-1} \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n}(x_i + y_i)^{(p-1)q}\right)^{1/q}$$

as well as

$$\sum_{i=1}^{n} y_i(x_i + y_i)^{p-1} \le \left(\sum_{i=1}^{n} y_i^p\right)^{1/p} \cdot \left(\sum_{i=1}^{n}(x_i + y_i)^{(p-1)q}\right)^{1/q}.$$

Summing up the last two inequalities and taking into account that $p + q = pq$, and so $(p - 1)q = pq - q = p$, we arrive at

$$\frac{\sum_{i=1}^{n}(x_i + y_i)^p}{\left(\sum_{i=1}^{n}(x_i + y_i)^p\right)^{1/q}} \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} + \left(\sum_{i=1}^{n} y_i^p\right)^{1/p}.$$

However, $1 - 1/q = 1/p$, so this is just the so-called *Minkowski inequality* which we have 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 $p \ge 1$. Beside that, we further set

$$\|z\|_\infty = \max\{|z_i|,\ i = 1, \dots, n\},$$

which, apparently, is a norm, too.

We can notice that Hölder's inequality can be, in the context of these norms, written for all $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ as

$$\sum_{i=1}^{n} |x_i| \cdot |y_i| \le \|x\|_p \cdot \|y\|_q$$

for all $p \ge 1$ and q satisfying $1/p + 1/q = 1$, where for $p = 1$, we set $q = \infty$.

7.16. L_p–norms for sequences and functions. Now we can easily define norms on suitable infinite-dimensional vector spaces as well. Let us begin with sequences. The vector space ℓ_p, $p \ge 1$, is the set of all sequences of real or complex numbers $x_0, x_1, \dots$ such that

$$\sum_{i=0}^{\infty} |x_i|^p < \infty.$$

All sequences with bounded absolute values of their terms create the space ℓ_∞. Taking the limit as n goes to ∞, we immediately get from the Minkowski inequality

$$\|x\|_p = \left(\sum_{i=0}^{\infty} |x_i|^p\right)^{1/p}$$

is a norm on ℓ_p. Similarly, we set

$$\|x\|_\infty = \sup\{|x_i|, i = 0, 1, \dots\}$$

on ℓ_∞, again obtaining a norm.

Thus, it makes sense to consider these four possibilities:

$$x \in X,\ x \notin Z,\ x \in Y, \quad x \in X,\ x \notin Z,\ x \notin Y,$$
$$x \notin X,\ x \in Z,\ x \in Y, \quad x \notin X,\ x \in Z,\ x \notin Y,$$

which may occur for $x \in (X \cup Z) \smallsetminus (X \cap Z)$. However, in any of these four cases, x belongs to exactly one of the sets $(X \cup Y) \smallsetminus (X \cap Y)$, $(Y \cup Z) \smallsetminus (Y \cap Z)$. Thus we have obtained the inclusion (||7.26||) whence the wanted triangle inequality follows.

$$d_1(X, Z) = |\,(X \cup Z) \smallsetminus (X \cap Z)\,| \leq$$
$$|\,[(X \cup Y) \smallsetminus (X \cap Y)] \cup [(Y \cup Z) \smallsetminus (Y \cap Z)]\,| \leq$$
$$|\,(X \cup Y) \smallsetminus (X \cap Y)\,| + |\,(Y \cup Z) \smallsetminus (Y \cap Z)\,| =$$
$$d_1(X, Y) + d_1(Y, Z).$$

The case (b). We can proceed similarly to the case of d_1. Let us denote by X' the complement of a set X. The equalities

$$(X \cup Y) \smallsetminus (X \cap Y) =$$
$$(X \cap Y' \cap Z) \cup (X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y \cap Z'),$$
$$(Y \cup Z) \smallsetminus (Y \cap Z) =$$
$$(X \cap Y \cap Z') \cup (X \cap Y' \cap Z) \cup (X' \cap Y \cap Z') \cup (X' \cap Y' \cap Z),$$
$$[(X \cup Z) \smallsetminus (X \cap Z)] \cup [Y \smallsetminus (X \cup Z)] =$$
$$(X \cap Y \cap Z') \cup (X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y' \cap Z) \cup (X' \cap Y \cap Z'),$$

which, again, can be proved by listing several possibilities, imply a stronger form of (||7.26||):

$$[(X \cup Z) \smallsetminus (X \cap Z)] \cup [Y \smallsetminus (X \cup Z)] \subseteq$$
$$[(X \cup Y) \smallsetminus (X \cap Y)] \cup [(Y \cup Z) \smallsetminus (Y \cap Z)].$$

Further, we invoke the inequality

$$\tfrac{|\,(X \cup Z) \smallsetminus (X \cap Z)\,|}{|\,X \cup Z\,|} \leq \tfrac{|\,[(X \cup Z) \smallsetminus (X \cap Z)] \cup [Y \smallsetminus (X \cup Z)]\,|}{|\,X \cup Z \cup [Y \smallsetminus (X \cup Z)]\,|}, \qquad X \cup Z \neq \emptyset.$$

That is based upon calculations with non-negative numbers only since it holds in general that

$$\tfrac{x}{z} \leq \tfrac{x+y}{z+y}, \qquad y \geq 0,\ z > 0,\ x \in [0, z].$$

Since, apparently,

$$X \cup Z \cup [Y \smallsetminus (X \cup Z)] = X \cup Y \cup Z,$$

we get

$$d_2(X, Z) = \tfrac{|\,(X \cup Z) \smallsetminus (X \cap Z)\,|}{|\,X \cup Z\,|} \leq \tfrac{|\,[(X \cup Z) \smallsetminus (X \cap Z)] \cup [Y \smallsetminus (X \cup Z)]\,|}{|\,X \cup Z \cup [Y \smallsetminus (X \cup Z)]\,|} \leq$$
$$\tfrac{|\,[(X \cup Y) \smallsetminus (X \cap Y)] \cup [(Y \cup Z) \smallsetminus (Y \cap Z)]\,|}{|\,X \cup Y \cup Z\,|} \leq \tfrac{|\,(X \cup Y) \smallsetminus (X \cap Y)\,| + |\,(Y \cup Z) \smallsetminus (Y \cap Z)\,|}{|\,X \cup Y \cup Z\,|} \leq$$
$$\tfrac{|\,(X \cup Y) \smallsetminus (X \cap Y)\,|}{|\,X \cup Y\,|} + \tfrac{|\,(Y \cup Z) \smallsetminus (Y \cap Z)\,|}{|\,Y \cup Z\,|} = d_2(X, Y) + d_2(Y, Z),$$

if $X \cup Z \neq \emptyset$ and $Y \neq \emptyset$. However, for $X = Z = \emptyset$ or $Y = \emptyset$, the triangle inequality clearly holds as well.

Therefore, both mappings are metrics. The metric d_1 has a mere helping use. On the other hand, the metric d_2 has wider applications and it is also known as Jaccard's metric. It is named after biologist Paul Jaccard, who, in 1908, described the measure of similarity in insects populations using the function $1 - d_2$. □

7.23. Let

$$d(x, y) := \tfrac{|\,x-y\,|}{1+|\,x-y\,|}, \qquad x, y \in \mathbb{R}.$$

Prove that d is a metric on $\mathbb{R}$.

Solution. Again, we prove the triangle inequality only (the rest is clear). Let us introduce a helping increasing function

$$f(t) := \frac{t}{1+t}, \qquad t \geq 0. \tag{7.27}$$

Eventually, let us get back to the space of functions $\mathcal{S}^0[a, b]$ on a finite interval $[a, b]$ or $\mathcal{S}^0_c[a, b]$ on an unbounded interval. We have already met the 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 so-called Riemann sums which correspond to splitting 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 $f(x)$ 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 just the product of the Riemann sums for the integrals $\|f\|_p$ and $\|g\|_q$.

Moving to limits, we thus verify the so-called *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 a compact support.

In just the same way as in the previous paragraph, we can derive the integral variant 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 will use the word "norm" for the entire space $S^0[a, b]$ 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 exceptional; we have realized it by the scalar 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 will always have suprema of absolute values on a finite closed interval, so we can set

$$\|f\|_\infty = \sup\{f(x),\ x \in [a, b]\}$$

The fact that f is increasing need not be verified by calculation of the first derivative. It can be seen by the simple rearrangement

$$f(s)-f(r)=\tfrac{s}{1+s}-\tfrac{r}{1+r}=\tfrac{s-r}{(1+s)(1+r)}>0,\qquad s>r\geq 0.$$

Therefore,

$$d(x,z)=\tfrac{|\,x-z\,|}{1+|\,x-z\,|}=\tfrac{|\,x-y+y-z\,|}{1+|\,x-y+y-z\,|}\leq\tfrac{|\,x-y\,|+|\,y-z\,|}{1+|\,x-y\,|+|\,y-z\,|}=$$
$$\tfrac{|\,x-y\,|}{1+|\,x-y\,|+|\,y-z\,|}+\tfrac{|\,y-z\,|}{1+|\,x-y\,|+|\,y-z\,|}\leq\tfrac{|\,x-y\,|}{1+|\,x-y\,|}+\tfrac{|\,y-z\,|}{1+|\,y-z\,|}=$$
$$d(x,y)+d(y,z),\qquad x,y,z\in\mathbb{R}.$$

□

7.24. Determine the distance of the functions

$$f(x)=x,\quad g(x)=-\tfrac{x}{\sqrt{1+x^2}},\quad x\in[1,2]$$

as elements of the normed vector space $\mathcal{S}[1,2]$ of continuous functions on the interval $[1,2]$ with norm

(a) $\|\,f\,\|_1=\int_1^2|\,f(x)\,|\,dx$;
(b) $\|\,f\,\|_\infty=\max\{|\,f(x)\,|;\ x\in[1,2]\}$.

Solution. The case (a). It suffices to compute

$$\int\limits_1^2|\,f(x)-g(x)\,|\ dx=\int\limits_1^2 x+\tfrac{x}{\sqrt{1+x^2}}\,dx=\left[\tfrac{x^2}{2}+\sqrt{1+x^2}\right]_1^2=$$
$$\tfrac{3}{2}+\sqrt{5}-\sqrt{2}.$$

The case (b). Now, we want to determine

$$\max_{x\in[1,2]}|\,f(x)-g(x)\,|=\max_{x\in[1,2]}\left(x+\tfrac{x}{\sqrt{1+x^2}}\right).$$

When looking for extrema of functions, differentiation is a very strong and efficient tool. From the inequality

$$\left(x+\tfrac{x}{\sqrt{1+x^2}}\right)'=1+\tfrac{1}{\left(\sqrt{1+x^2}\right)^3}>0,\qquad x\in[1,2],$$

we can immediately see

$$\max_{x\in[1,2]}\left(x+\tfrac{x}{\sqrt{1+x^2}}\right)=2+\tfrac{2}{\sqrt{1+2^2}}=2+\tfrac{2}{\sqrt{5}}.$$

An increasing function takes the maximum value at the right marginal point of a closed interval. □

7.25. Determine whether the sequence $\{x_n\}_{n\in\mathbb{N}}$ where

$$x_1=1,\qquad x_n=1+\tfrac{1}{2}+\cdots+\tfrac{1}{n},\ n\in\mathbb{N}\smallsetminus\{1\},$$

is a Cauchy sequence in $\mathbb{R}$. First, consider the usual metric given by the difference in absolute value (i. e., induced by the norm of absolute value). Then, consider the metric

$$d(x,y):=\tfrac{|\,x-y\,|}{1+|\,x-y\,|},\qquad x,y\in\mathbb{R}.$$

Solution. Let us remind that

$$\sum_{k=1}^{\infty}\frac{1}{k}=\infty,\quad\text{i. e.}\quad\sum_{k=m}^{\infty}\frac{1}{k}=\infty,\ m\in\mathbb{N}.\tag{7.28}$$

Therefore,

$$\lim_{n\to\infty}|\,x_n-x_m\,|=\sum_{k=m+1}^{\infty}\tfrac{1}{k}=\infty,\qquad m\in\mathbb{N}.$$

for such a function f. Let us notice that 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 the values at discontinuity points).

7.17. 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. Actually, this is contained in the axiom of existence of suprema. Let us remind the the real numbers were created as a "completion" of the space of rational numbers which is not complete itself. It is apparent that the closure of the set $\mathbb{Q}\subset\mathbb{R}$ is the whole $\mathbb{R}$.

Dense and nowhere-dense subsets

We say that a subset $A\subset X$ in a metric space X is *dense* iff the closure of A is the whole space X. A set A is said to be *nowhere dense* in X iff the set $X\setminus\bar{A}$ is dense.

Apparently, 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, we can easily see that 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. Let us 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 $[1/n,1]$. Apparently, they converge to the function $\sin(1/x)$ in all L_p norms, but this function does not lie in our spaces.

Completion of a metric space

Let X be a metric space with metric d which is not complete. A metric space $\tilde{X}$ with metric $\tilde{d}$ such that $X\subset\tilde{X}$, d is the restriction of $\tilde{d}$ to the subset X and the closure $\bar{X}$ 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. Before we get to the quite difficult proof of this extraordinarily important and useful result, we can notice that such a "completion" $\tilde{X}$ of a space X can be done in a unique way, in a certain sense:

A mapping $\varphi:X_1\to X_2$ between metric spaces with metrics d_1 and d_2, respectively, is called an *isometry* iff 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 of distinct elements is non-zero) and the corresponding inverse mapping is an isometry as well.

Now, let us 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 let us denote the corresponding metrics by d, d_1, and d_2, respectively. Apparently, the mapping

$$\varphi:\ \iota_1(X)\xrightarrow{\iota_1^{-1}}X\xrightarrow{\iota_2}\tilde{X}_2$$

Hence we can see that the sequence $\{x_n\}$ cannot be a Cauchy sequence. Thus we have found the answer for the usual metric. However, we could have utilized the fact that the sequence $\{x_n\}$ is not convergent by (‖7.28‖) and that we find ourselves in a complete metric space, where Cauchy sequences and convergent sequences coincide.

For the metric d, it suffices to realize that the mapping f introduced in (‖7.27‖) is a continuous bijection between the sets $[0, \infty)$ and $[0, 1)$, having the property that $f(0) = 0$. Thus, any sequence is convergent "in the original meaning" if and only if it converges in the metric space $\mathbb{R}$ with metric d. It holds as well that a sequence is a Cauchy sequence in $\mathbb{R}$ with respect to the usual metric if and only if it is a Cauchy sequence with respect to d. □

7.26. Is the metric space $\mathcal{S}[-1, 1]$ of continuous functions on the interval $[-1, 1]$ with metric given by the norm

$$\text{(a)}\quad \| f \|_p = \left(\int_{-1}^{1} | f(x) |^p \, dx \right)^{1/p} \quad \text{for } p \geq 1;$$

$$\text{(b)}\quad \| f \|_\infty = \max \{ | f(x) |;\ x \in [-1, 1] \}$$

complete?

Solution. The case (a). Let us, for every $n \in \mathbb{N}$, define a function

$$f_n(x) = 0,\ x \in [-1, 0)\,, \qquad f_n(x) = 1,\ x \in \left[\tfrac{1}{n}, 1\right],$$
$$f_n(x) = nx,\ x \in \left[0, \tfrac{1}{n}\right).$$

The obtained sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{S}[-1, 1]$ is a Cauchy sequence of functions. To verify it is a Cauchy sequence, it suffices, using the geometrical meaning of definite integral, to express

$$\left(\int_{-1}^{1} | f_m(x) - f_n(x) |^p \, dx \right)^{1/p} < \left(\int_{0}^{1/n} 1 \, dx \right)^{1/p} = \left(\tfrac{1}{n}\right)^{1/p}$$

for every $m \geq n$, $m, n \in \mathbb{N}$.

Let us focus on the potential limit of the sequence $\{f_n\}$ in $\mathcal{S}[-1, 1]$. Let us assume it exists and denote it by f. For every $\varepsilon \in (0, 1)$, there apparently exists an $n(\varepsilon) \in \mathbb{N}$ such that

$$f_n(x) = 0,\ x \in [-1, 0], \qquad f_n(x) = 1,\ x \in [\varepsilon, 1]$$

for all $n \geq n(\varepsilon)$. Therefore, the continuous function f must satisfy

$$f(x) = 0,\ x \in [-1, 0], \qquad f(x) = 1,\ x \in [\varepsilon, 1]$$

for an arbitrarily small $\varepsilon > 0$. Thus, necessarily,

$$f(x) = 0,\ x \in [-1, 0], \qquad f(x) = 1,\ x \in (0, 1].$$

However, this function is not continuous on $[-1, 1]$ – it does not belong to the considered metric space. Therefore, the sequence $\{f_n\}$ does not have a limit in $\mathcal{S}[-1, 1]$, so this space is not complete.

The case (b). Let an arbitrary Cauchy sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{S}[-1, 1]$ 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

$$\text{(7.29)} \qquad \max_{x\in[-1,1]} | f_m(x) - f_n(x) | < \frac{\varepsilon}{2}, \qquad m, n \geq n(\varepsilon).$$

In particular, we get for every $x \in [-1, 1]$ 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. Let us set

$$f(x) := \lim_{n\to\infty} f_n(x), \qquad x \in [-1, 1].$$

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 surely 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 the previous reasoning, 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.

7.18. Theorem. *Let X be a metric space with metric d which is not complete. Then there exists a completion $\tilde{X}$ of X with metric $\tilde{d}$ which is unique up to bijective isometries.*

Proof. The idea of the construction is quite 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 iff $d(x_i, y_i)$ 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 quite apparent that the relation defined above is really 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.

Now, let us define $\tilde{X}$ as the set of the classes of this equivalence of Cauchy sequences. The original points $x \in X$ can be identified with the class of sequences equivalent to the constant sequence $x_i = x$, $i = 0, 1, \dots$.

It is now easy to define the metric $\tilde{d}$. It suggests itself to consider

$$\tilde{d}(\tilde{x}, \tilde{y}) = \lim_{i\to\infty} d(x_i, y_i)$$

for sequences $\tilde{x} = \{x_0, x_1, \dots\}$ and $\tilde{y} = \{y_0, y_1, \dots\}$.

First, we have to verify that this limit exists at all and is finite. Straight from 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 $\tilde{x} = \{x'_0, x'_1, \dots\}$ and $\tilde{y} = \{y'_0, y'_1, \dots\}$, then we can see from the triangle inequality for the distance of real numbers (we need to consider the consequences for differences of distances) that

$$\begin{aligned} |d(x'_i, y'_i) - d(x_i, y_i)| &\leq |d(x'_i, y'_i) - d(x'_i, y_i)| + \\ &\qquad |d(x'_i, y_i) - d(x_i, y_i)| \\ &\leq d(x_i, x'_i) + d(y_i, y'_i). \end{aligned}$$

Therefore, the definition is indeed independent of the choice of representatives.

Further, 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, let us choose three Cauchy representatives of the

Letting $m \to \infty$ in (‖7.29‖), we obtain

$$\max_{x\in[-1,1]} |\, f(x) - f_n(x)\,| \le \tfrac{\varepsilon}{2} < \varepsilon, \quad n \ge n(\varepsilon).$$

However, this means that the sequence $\{f_n\}_{n\in\mathbb{N}}$ converges uniformly to the function f on $[-1, 1]$. In other words, $\{f_n\}_{n\in\mathbb{N}}$ converges to f with respect to the given norm. We have already found that a uniform limit of continuous functions is a continuous function. Thanks to that, we need not prove that $f \in \mathcal{S}[-1, 1]$. Therefore, the metric space is complete.

Let us mention that we could arrive at the same results (using the same reasoning in both cases) for the more general metric space $\mathcal{S}[a, b]$ of continuous functions on $[a, b]$ as well. □

7.27. One of the most important classifications of metric spaces is given by the so-called principle of nested balls. 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 satisfies

$$\bigcap_{n\in\mathbb{N}} A_n \neq \emptyset. \tag{7.30}$$

However, there is one more part of this theorem - a requirement upon the considered sequences $\{A_n\}$. It must hold that

$$\lim_{n\to\infty} \sup\,\{d(x, y);\ x, y \in A_n\} = 0. \tag{7.31}$$

Find out whether this requirement can be omitted.

Solution. The requirement (‖7.31‖), probably contrarily to many readers' expectations, cannot be omitted: the statement would become false. We need to give a counterexample proving that the statement does not hold without the mentioned condition.

For that purpose, let us consider the set $X = \mathbb{N}$ with metric

$$d(m, n) = 1 + \tfrac{1}{m+n}, \ \ m \neq n, \quad d(m, n) = 0, \ \ m = n.$$

The first and second properties are clearly satisfied. To prove the triangle inequality, it suffices to realize that $d(m, n) \in (1, 4/3]$ if $m \neq n$. All Cauchy sequences can be found equally easily: they are the so-called almost stationary sequences — constant from some index on (i. e., constant except for finitely many terms). Thus, every Cauchy sequence is convergent, so the metric space in question is complete.

Let us introduce the sets

$$A_n := \left\{m \in \mathbb{N};\ d(m, n) \le 1 + \tfrac{1}{2n}\right\}, \quad n \in \mathbb{N}.$$

As the inequality in their definition is not strict, it is guaranteed that they are closed sets. Since $A_n = \{n, n + 1, \dots\}$, (‖7.30‖) does not hold. If we omitted the requirement (‖7.31‖), it would mean that the metric space is not complete, which is not true. Finally, let us mention that

$$\lim_{n\to\infty} \sup\,\{d(x, y);\ x, y \in A_n\} = \lim_{n\to\infty} \left(1 + \tfrac{1}{2n+1}\right) = 1 \neq 0.$$

□

7.28. Prove that the metric space l_2 is complete.

Solution. Let us consider an arbitrary Cauchy sequence $\{x_n\}_{n\in\mathbb{N}}$ in the space l_2. However, every term of this sequence is again a sequence, i. e., $x_n = \{x_n^k\}_{k\in\mathbb{N}}$, $n \in \mathbb{N}$. Let us mention that, of course, the range of indeces does not matter – there is no difference whether $n, k \in \mathbb{N}$ or

elements $\tilde{x}$, $\tilde{y}$, $\tilde{z}$, and we easily get

$$\begin{aligned}\tilde{d}(\tilde{x}, \tilde{z}) &= \lim_{i\to\infty} d(x_i, z_i)\\ &\le \lim_{i\to\infty} d(x_i, y_i) + \lim_{i\to\infty} d(y_i, z_i)\\ &= \tilde{d}(\tilde{x}, \tilde{y}) + \tilde{d}(\tilde{y}, \tilde{z}).\end{aligned}$$

Apparently, 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 remains to prove that X is dense in $\tilde{X}$ and that the constructed metric space is complete. We want to prove that for any fixed Cauchy sequence $\tilde{x} = \{x_i\}$ and every (no matter how small) $\varepsilon > 0$, we can find an element y of the original space such that the distance of the constant sequences of y's from the chosen sequence x_i does not exceed ε. However, since the sequence x_i is a Cauchy sequence, all pairs of its terms x_n, x_m will eventually (i. e. for sufficiently large indeces m and n) become closer than ε to each other. Then the choice $y = x_n$ for one of those indeces necessarily gives that the elements y and x_m will be closer than ε, and so, from the limit point of view, it will hold that $\tilde{d}(\tilde{y}, \tilde{x}) \le \varepsilon$.

Finally, it remains to prove that Cauchy sequences of points of the extended space $\tilde{X}$ with respect to the metric $\tilde{d}$ are necessarily convergent. In other words, we want to show that repeating the above procedure does not yield new points. This can be done by approaching 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 we already know that 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, let us consider the sequence $\tilde{z} = \{z_0, z_1, \dots\}$. The original sequence $\tilde{x}$ is Cauchy, i. e. 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, we can assume that our index $n(\varepsilon)$ is greater than or equal to $4/\varepsilon$. Now, for m and n greater than $n(\varepsilon)$, we get:

$$\begin{aligned}d(z_m, z_n) &= \tilde{d}(\tilde{z}_m, \tilde{z}_n)\\ &\le \tilde{d}(\tilde{z}_m, \tilde{x}_m) + \tilde{d}(\tilde{x}_m, \tilde{x}_n) + \tilde{d}(\tilde{x}_n, \tilde{z}_n)\\ &\le 1/m + \varepsilon/2 + 1/n \le 2\frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Thus it is a Cauchy sequence z_i of elements in X, and so $\tilde{z} \in \tilde{X}$. Let us examine whether the distance $\tilde{d}(\tilde{x}_n, \tilde{z})$ approaches zero, which we tried to guarantee by the construction. From the triangle inequality,

$$\tilde{d}(\tilde{z}, \tilde{x}_n) \le \tilde{d}(\tilde{z}, \tilde{z}_n) + \tilde{d}(\tilde{z}_n, \tilde{x}_n).$$

However, from our previous bounds, it follows that both the summands on the right-hand side converge to zero, thereby finishing the proof. □

In the following three paragraphs, we will introduce quite simple three theorems about complete metric spaces. They are highly applicable in both mathematical analysis and verifying convergence of numerical methods.

$n, k \in \mathbb{N} \cup \{0\}$. Let us introduce helping sequences y_k for $k \in \mathbb{N}$ so that

$$y_k = \{y_k^n\}_{n\in\mathbb{N}} = \left\{x_n^k\right\}_{n\in\mathbb{N}}.$$

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. Let us denote their limits by z_k, $k \in \mathbb{N}$.

It suffices to prove that $z = \{z_k\}_{k\in\mathbb{N}} \in l_2$ and that the sequence $\{x_n\}$ converges for $n \to \infty$ in l_2 just to the sequence z. The sequence $\{x_n\}_{n\in\mathbb{N}} \subset l_2$ 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 $m \to \infty$, we can obtain

$$\sum_{k=1}^{l} \left(z_k - x_n^k\right)^2 \leq \varepsilon, \quad n \geq n(\varepsilon),\ n, l \in \mathbb{N},$$

i. e. (this time $l \to \infty$)

$$\sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 \leq \varepsilon, \quad n \geq n(\varepsilon),\ n \in \mathbb{N}. \tag{7.32}$$

Especially, we have

$$\sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 < \infty, \quad n \geq n(\varepsilon),\ n \in \mathbb{N}$$

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 $\{x_n\}_{n\in\mathbb{N}} \subset l_2$. Since

$$\sum_{k=1}^{\infty} \left(z_k x_n^k\right) \leq \sqrt{\sum_{k=1}^{\infty} z_k^2} \cdot \sqrt{\sum_{k=1}^{\infty} \left(x_n^k\right)^2}, \quad n \in \mathbb{N}$$

and

$$\sum_{k=1}^{\infty} \left(z_k - x_n^k\right)^2 = \sum_{k=1}^{\infty} \left[z_k^2 - 2z_k x_n^k + \left(x_n^k\right)^2\right], \quad n \in \mathbb{N},$$

it must be that

$$\sum_{k=1}^{\infty} z_k^2 < \infty.$$

Thus we have proved that $z \in l_2$. The fact that $\{x_n\}$ converges for $n \to \infty$ to z in l_2 follows from (‖7.32‖). □

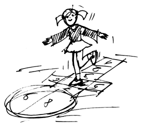
7.29. In the metric space $\mathcal{S}[-1, 1]$ with metric given by the norm $\|\cdot\|_\infty$, consider the sets

$$A = \{f \in \mathcal{S}[-1, 1];\ f(0) \in (0, 2)\},$$

$$B = \{f \in \mathcal{S}[-1, 1];\ \int_{-1}^{1} f(x)\,dx = 0\}.$$

Are these sets open, closed?

7.19. Banach's contraction principle. A mapping $F : X \to X$ on a metric space X with metric d is called a *contraction mapping* iff 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 suggesting that if iterative application of a contraction mapping to an initial value $z_0 \in X$ should "accumulate" to some point. The metric space X, of course, needs to be complete; otherwise it could happen that the limit point does not exist in it.

Let us 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

$$\begin{aligned} d(z_{i+1}, z_i) &= d(F(z_i), F(z_{i-1})) \\ &\leq C\, d(z_i, z_{i-1}) \leq \cdots \leq C^i\, d(z_1, z_0). \end{aligned}$$

The triangle inequality then implies that for all natural numbers j,

$$\begin{aligned} d(z_{i+j}, z_i) &\leq \sum_{k=1}^{j} d(z_{i+k}, z_{i+k-1}) \\ &\leq \sum_{k=1}^{j} C^{i+k-1} d(z_1, z_0) = C^i\, d(z_1, z_0) \sum_{k=1}^{j} C^{k-1} \\ &\leq C^i\, d(z_1, z_0) \sum_{k=1}^{\infty} C^{k-1} = \frac{C^i}{1 - C}\, d(z_1, z_0). \end{aligned}$$

Now, for every positive (no matter how small) ε, the right-hand expression is surely less than ε for sufficiently large indeces i, i. e.,

$$d(z_i, z_{i+j}) \leq \frac{C^i}{1 - C}\, d(z_1, z_0) \leq \varepsilon.$$

However, this just says that our sequence z_i is a Cauchy sequence. Thanks to the space X being complete, its limit z surely exists, and all that remains to be proved is $F(z) = z$.

However, every contraction mapping is clearly continuous. Therefore,

$$F(z) = F(\lim_{n\to\infty} z_n) = \lim_{n\to\infty} F(z_n) = z.$$

This finishes the proof. □

7.20. 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* iff $\operatorname{diam} A < \infty$.

Theorem. *If $A_1 \supset A_2 \supset \cdots \supset A_i \supset \ldots$ is a non-decreasing sequence of non-empty closed subsets in a complete metric space X and* $\operatorname{diam} A_i \to 0$*, then there is exactly one point $x \in X$ belonging to the intersection of all the sets A_i.*

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 $\overline{M}$. We can easily see that a set M is closed if and only if $M = \overline{M}$. Since

$$A^0 = A,\ \overline{A} = \{f \in \mathcal{S}[-1, 1];\ f(0) \in [0, 2]\}, \quad B^0 = \emptyset,\ \overline{B} = B,$$

the set A is open and not closed, and, the other way around, the set B is closed and not open. □

7.30. Let an arbitrary set $X \neq \emptyset$ be given. The mapping $d : X \times X \to \mathbb{R}$ defined by the formula

$$d(x, y) := 1,\ x \neq y, \quad d(x, y) := 0,\ x = y$$

is clearly a metric on X. We talk about the so-called trivial or (more often) discrete metric space (X, d).

(a) Describe all Cauchy and convergent sequences in (X, d).
(b) Describe all open, closed, and bounded sets in (X, d).
(c) Describe 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). 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 a common value $x_n = x_{n+1} = \cdots$ (we talk about almost stationary sequences). Besides others, we have proved that the metric space (X, d) is complete.

The case (b). The open 1-neighborhood 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). However, this also means that every set is closed as well. The fact that the 2-neighborhood of any element coincides with the whole space implies that every set is bounded in (X, d).

The case (c). Once again, we make use of the fact that the open 1-neighborhood of any element contains this element only. Hence 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 apparently 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). □

7.31. Decide whether the set (known as Hilbert cube)

$$A = \left\{\{x_n\}_{n\in\mathbb{N}} \in l_2;\ |\,x_n\,| \le \tfrac{1}{n},\ n \in \mathbb{N}\right\}$$

is compact in l_2. Then, decide the compactness of the set

$$B = \left\{\{x_n\}_{n\in\mathbb{N}} \in l_\infty;\ |\,x_n\,| < \tfrac{1}{n},\ n \in \mathbb{N}\right\}$$

in the space l_∞.

Solution. We know that the space l_2 is complete. Every closed subset of a complete metric space defines a complete metric space. The set A is apparently closed in l_2, so it suffices to show that it is totally bounded, and we will get its compactness.

Let us begin with the well-known series

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Proof. Let us select one point z_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 indeces $i \ge n(\varepsilon)$, their diameters are less than ε. However, then for so large indeces i, j, we will have $d(z_i, z_j) \le \varepsilon$, and thus our sequence is a Cauchy sequence. Therefore, it has a limit point $z \in X$, which, of course, must be a limit point of all the sets A_i, thus it belongs to all of them (since they are all closed) and so to their intersection.

We have proved the existence of z. Now, it remains to prove its uniqueness. For that purpose, assume there are points z and y, both belonging to the intersection of all the sets A_i. Then their distance must be less than the diameter of the sets A_i, but that converges to zero. This completes the proof. □

7.21. 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.*

Proof. Let a system of dense open sets A_i, $i = 1, 2\ldots$, be given in X. We want to show that the set $A = \cap_{i=1}^{\infty} A_i$ has a non-empty intersection with any open set $U \subset X$. We will proceed inductively, invoking the previous theorem.

Surely there is a $z_1 \in A_1 \cap U$, but since the set A_1 is open, the closure of an ε_1-neighborhood U_1 (for sufficiently small ε_1) of the point z_1 is contained in A_1 as well. Let us denote the closure of this ε_1–ball U_1 by B_1. Further, let us suppose that the points z_i and their open ε_i–neighborhoods 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}-neighborhood 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 $\varepsilon_n \le 1/n$.

If we proceed in this inductive way from the original point z_1 and the set B_1, we get a non-decreasing sequence of non-empty closed sets B_n whose diameter approaches zero. Therefore, there is a point z common to all of these sets, i. e.,

$$z \in \cap_{i=1}^{\infty} \bar{U}_i = \cap_{i=1}^{\infty} B_i \subset \cap_{i=1}^{\infty} A_n \cap U,$$

which is the statement to be proved. □

7.22. Bounded and compact sets. The following concepts facilitated the phrasing of our observations about the real numbers. They can be reformulated for general metric spaces with almost no changes.:

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–neighborhoods.

A *boundary point* of a set A is an element $x \in X$ such that each of its neighborhoods 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.

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 the whole of A.

An *isolated point* of a set A is an element $a \in A$ such that one of its ε–neighborhoods in X has the singleton intersection $\{a\}$ with A.

A set A of elements of a metric space is called *bounded* iff its diameter is finite, i. e., there is a real number r such that $d(x, y) \le$

Therefore, for every $\varepsilon > 0$, there is an $n(\varepsilon) \in \mathbb{N}$ satisfying

$$\sqrt{\sum_{k=n(\varepsilon)+1}^{\infty} \frac{1}{k^2}} < \frac{\varepsilon}{2}.$$

>From each of the intervals $[-1/n, 1/n]$ for $n \in \{1, \dots, n(\varepsilon)\}$, we can choose finitely many points $x_1^n, \dots, x_{m(n)}^n$ so that we would have for any $x \in [-1/n, 1/n]$ that

$$\min_{j \in \{1,\dots,m(n)\}} \left| x - x_j^n \right| < \frac{\varepsilon}{\sqrt{5^n}}.$$

Let us consider such sequences $\{y_n\}_{n\in\mathbb{N}}$ from l_2 whose terms with indeces $n > n(\varepsilon)$ are zero, and at the same time,

$$y_1 \in \left\{x_1^1, \dots, x_{m(1)}^1\right\}, \dots, y_{n(\varepsilon)} \in \left\{x_1^{n(\varepsilon)}, \dots, x_{m(n(\varepsilon))}^{n(\varepsilon)}\right\}.$$

There are only finitely many such sequences and they create an ε-net for A since

$$\sqrt{\tfrac{\varepsilon^2}{5} + \tfrac{\varepsilon^2}{5^2} + \cdots + \tfrac{\varepsilon^2}{5^{n(\varepsilon)}}} + \tfrac{\varepsilon}{2} < \varepsilon \cdot \sqrt{\tfrac{1}{1-\frac{1}{5}} - 1} + \tfrac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the set A is totally bounded, which implies its compactness.

It is very simple to determine whether the set B is compact. Every compact set must be closed, but the set B is not. Its closure is

$$\overline{B} = \left\{\{x_n\}_{n\in\mathbb{N}} \in l_\infty;\ |x_n| \le \tfrac{1}{n},\ n \in \mathbb{N}\right\}.$$

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. □

D. Integral operators

The convolution is one of the tools for smoothing functions:

7.32. Determine the convolution $f_1 * f_2$ where

$$\begin{aligned} f_1(x) &= \frac{1}{x} \quad \text{for } x \neq 0 \\ f_2(x) &= \begin{cases} x & \text{for } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Solution. The value of the convolution at a point t is given by the integral $\int_{-\infty}^{\infty} f_1(x) f_2(t-x)\, \mathrm{d}x$. The integrated function is non-zero if the second factor is non-zero, i. e., if $(t-x) \in [-1, 1]$, i. e., $x \in [t-1, t+1]$. The value of the convolution at the point t can be interpreted as the integral mean of the function f_1 over the interval $(t-1, t+1)$. When integrating over this interval, we have to distinguish whether the number 0 belongs to it. If the interval contains zero, the integral must be split into two improper integrals. However, the value of the smaller one can be subtracted thanks to the function $\frac{1}{x}$ being odd, so the integral $\int_{1-t}^{t+1} \frac{1}{x}\, \mathrm{d}x$ remains (think out why the formula works for negative numbers t as well). Thus, we get:

$$f_1 * f_2(t) = \left\{ \begin{array}{ll} \int_{t-1}^{t+1} \frac{1}{x}\, \mathrm{d}x = \ln\left|\frac{t+1}{t-1}\right| & \text{for } t \in (-\infty, -1] \cup [1, \infty], \\ \int_{1-t}^{1+t} \frac{1}{x}\, \mathrm{d}x = \ln\left|\frac{1+t}{1-t}\right| & \text{for } t \in [-1, 1]. \end{array} \right.$$

□

Now, let us try to calculate the convolution of two functions both of which have a finite support.

r for all elements $x, y \in A$. In the other case, the set is said to be *unbounded.*

A metric space X is called *compact* iff every sequence of terms $x_i \in X$ has a subsequence converging to some point $x \in X$.

In the case of the real numbers, we mentioned several characterizations of compactness. The concept of boundedness is a bit more complicated in the case of metric spaces. For any subsets $A, B \subset X$ in a metric space X with metric d, we define their *distance*

$$\operatorname{dist}(A, B) = \sup_{x\in A, y\in B} \{d(x, y)\}.$$

If $A = \{x\}$ is a singleton set, we talk about the distance $\operatorname{dist}(x, B)$ of the point x from the set B. We say that a metric space X is *totally bounded* iff for every positive real number ε, there is a finite set A such that

$$\operatorname{dist}(x, A) < \varepsilon$$

for all points $x \in X$. Let us remind that a metric space is *bounded* iff the whole X has a finite diameter.

We can immediately see that a totally bounded space is especially bounded. Indeed, the diameter of a finite set is always 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 $\operatorname{dist}(x, A)$, $\operatorname{dist}(y, A)$, and $\operatorname{diam} A$, 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 the total boundedness. (Verify this in detail by yourselves!)

Theorem. *The following statements about a metric space X are equivalent:*

(1) X is compact,
(2) every open cover of X contains a finite cover,
(3) X complete and totally bounded.

SKETCH OF THE PROOF. If the second statement of the theorem is satisfied, the we can easily see that the space X must be totally bounded. Indeed, it suffices to choose the cover of X consisting of all ε–balls centered at the points $x \in X$. We can choose a finite cover from that and the set of centers x_i of the balls participating in this finite cover already satisfies the condition from the definition of total boundedness.

To prove the implication (2) $\Longrightarrow$ (3), we need to show the completeness. Let us consider a Cauchy sequence x_i .

□

7.23. Compactness on continuous functions. As an example of the fact that the behavior of compactness may differ in Euclidean spaces from that in spaces of functions, we will mention a very useful theorem, known as *Arzela-Askoli theorem.*

Theorem. *A set $M \subset \mathcal{C}[a, b]$ is compact if and only if it is bounded, closed, and equicontinuous.*

7.33. 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. The value of the convolution $f_1 * f_2$ at a point t is given by the integral over all real numbers of the product of the function $f_1(x)$ and the function $f_2(t-x)$ with respect to the variable x (see 7.13). Thus, this value is zero if either of the values $f_1(x)$ and $f_2(t-x)$ is zero for any real x. On the other hand, the value of the convolution can be non-zero at a point t only if there are numbers x such that $f_1(x) \neq 0 \neq f_2(t-x)$. By the definitions of the given functions, this occurs if there are numbers $x \in [-1,1]$ ($f_1(x) \neq 0$) such that $(t-x) \in [0,1]$ ($f_2(t-x) \neq 0$). I. e., $f_1 * f_2(t)$ can be non-zero if $[t-1, t+1] \cap [0,1] \neq \emptyset$. This happens for $t \in [-1,2]$. We integrate over x belonging to the intersection of the intervals $[t-1, t+1]$ and $[0,1]$. Further, this intersection depends on $t \in [-1,2]$:

a) for $t \in [-1,0]$, we have $[t-1, t+1] \cap [0,1] = [0, t+1]$,
b) for $t \in [0,1]$, we have $[t-1, t+1] \cap [0,1] = [0,1]$,
c) for $t \in [1,2]$, we have $[t-1, t+1] \cap [0,1] = [t-1, 1]$.

According to the intersection of these intervals, we then have: a)

$$\int_{-\infty}^{\infty} f_1(x) f_2(t-x)\,\mathrm{d}x = \int_0^{t+1} f_1(x) f_2(t-x)\,\mathrm{d}x$$
$$= \int_0^{t+1} (1-x^2)(t-x)\,\mathrm{d}x = -\frac{1}{4}t^4 + t^2 + \frac{2}{3}t - \frac{1}{4},$$

b)

$$\int_{-\infty}^{\infty} f_1(x) f_2(t-x) = \int_0^1 f_1(x) f_2(t-x)$$
$$= \int_0^1 (1-x^2)(t-x)\,\mathrm{d}x = \frac{2}{3}t - \frac{1}{4},$$

c)

$$\int_{-\infty}^{\infty} f_1(x) f_2(t-x) = \int_{t-1}^1 f_1(x) f_2(t-x)$$
$$= \int_{t-1}^1 (1-x^2)(t-x)\,\mathrm{d}x = \frac{1}{12}t^4 - t^2 + \frac{4}{3}t.$$

Altogether, we get:

$$f_1 * f_2(t) = \begin{cases} -\frac{1}{4}t^4 + t^2 + \frac{2}{3}t - \frac{1}{4} & \text{for } t \in [-2,-1], \\ \frac{2}{3}t - \frac{1}{4} & \text{for } t \in [-1,1], \\ \frac{1}{12}t^4 - t^2 + \frac{4}{3}t & \text{for } t \in [1,2] \\ 0 & \text{otherwise.} \end{cases}$$

□

7.24. Proof of theorem 7.8 about Fourier series. The general context of metrics and convergence allows us now to get back to the proof of the theorem in which we got a first idea about piecewise and other convergences of Fourier series. However, we do not care about necessary conditions for convergence, and many other formulations can be found in literature. On the other hand, our theorem 7.8 was quite simple and concerned a good deal of useful cases.

Firstly, it is good to realize how convergences may differ with respect to different L_p norms. For the sake of simplicity, we will always work in the completion of the space S_c^0 or S_c^1 with respect to the corresponding norm, without thinking about what the spaces actually look like (even though they could be described quite easily with the help of Kurzweil integral).

Hölder's inequality (applied to functions f and constant 1) yields the first of the following bounds on $\mathcal{S}^0[a,b]$:

$$\int_a^b |f(x)| dx \le |a-b|^{1/q} \left(\int_a^b |f(x)|^p\, dx \right)^{1/p}$$
$$\le |a-b|^{1/q} C^{1/q} \left(\int_a^b |f(x)|\, dx \right)^{1/p},$$

where $p > 1$ and $1/p + 1/q = 1$, $C \ge |f(x)|$ on the whole interval $[a,b]$ (such a uniform bound by a constant always exists if $f \in \mathcal{S}^0[a,b]$). The second bound follows immediately from the bound $|f(x)|^p \le C^{p-1}|f(x)|$ and the relation $1 - 1/p = 1/q$.

Thus it is apparent from the first bound that L_p–convergence $f_n \to f$ is, for any $p > 1$, always stronger than L_1–convergence (and with a merely modified bound, we can derive an even stronger proposition, namely that L_q–convergence is stronger than L_p–convergence whenever $q > p$; try this by yourselves). However, to apply the second bound, we have to require uniform boundedness of the sequence of functions f_n, i. e. the bound for the functions f_n by a constant C must be independent of n. Then we can assert that $|f_n(x) - f(x)| \le 2C$, and our bound implies that L_1–convergence is stronger than L_p–convergence.

Therefore, all our L_p–norms on our space $\mathcal{S}^0[a,b]$ are equivalent with regard to convergence of uniformly bounded sequences of functions.

The most difficult (and most interesting) part is to prove the first statement of the theorem 7.8, which is in literature often referenced as *Dirichlet condition* (it is deemed to be derived as early as in 1824). First, we prove how this property of piecewise convergence implies the statements (2) and (3) of the theorem to be proved. Without loss of generality, we can assume that we are working on the interval $[-\pi, \pi]$, i. e. with period $T = 2\pi$.

As the first step, we prepare simple bounds for the coefficients of the Fourier series. A bound-of-course is

$$|a_n| \le \frac{1}{\pi} \int_{-\pi}^{\pi} |f(x)|\, dx,$$

and the same for all the coefficients b_n since 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

$$a_n(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\cos(nx) dx$$
$$= \frac{1}{n\pi}[f(x)\sin(nx)]_{-\pi}^{\pi} - \frac{1}{n\pi} \int_{-\pi}^{\pi} f'(x)\sin(nx)\, dx$$

7.34. Determine the convolution $f_1 * f_2$ of the functions

$$f_1 = \begin{cases} 1 - x & \text{for } x \in [-2, 1], \\ 0 & \text{otherwise}, \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{for } x \in [0, 1], \\ 0 & \text{otherwise}. \end{cases}$$

○

7.35. Find the Fourier transform $\mathcal{F}(f) = \tilde{f}$ of the function

$$f(t) = \operatorname{sgn} t, \quad t \in (-1, 1)\,; \qquad f(t) = 0, \quad t \in \mathbb{R} \smallsetminus (-1, 1)\,,$$

i. e., $f(0) = 0$, $f(t) = 1$ for $t \in (0, 1)$ and $f(t) = -1$ for $t \in (-1, 0)$.

Solution. The Fourier transform of the given function is

$$\mathcal{F}(f)(\omega) = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-i\omega t}\, dt =$$

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-1}^{1} \operatorname{sgn} t\, \left[\cos(\omega t) - i \sin(\omega t)\right]\, dt.$$

Since the product of any two odd functions is an even function, the product of an even function and an odd function is an odd function and since the integral of an odd function over the interval $[-1, 1]$ is 0 (if this integral exists at all) and the integral of an even function over the interval $[-1, 1]$ is twice the integral over $[0, 1]$, we further get

$$\mathcal{F}(f)(\omega) = \tfrac{2}{\sqrt{2\pi}} \int\limits_{0}^{1} -i \sin(\omega t)\; dt = \tfrac{2i}{\sqrt{2\pi}} \left[\tfrac{\cos(\omega t)}{\omega}\right]_0^1 = i\sqrt{\tfrac{2}{\pi}}\, \tfrac{\cos\omega - 1}{\omega}.$$

If we directly used the known expression of the Fourier transform of an odd function f, we would obtain more easily that

$$\mathcal{F}(f)(\omega) = \tfrac{-2i}{\sqrt{2\pi}} \int\limits_{0}^{\infty} f(t)\, \sin(\omega t)\; dt = \tfrac{-2i}{\sqrt{2\pi}} \int\limits_{0}^{1} \sin(\omega t)\; dt = \cdots =$$

$$i\sqrt{\tfrac{2}{\pi}}\, \tfrac{\cos\omega - 1}{\omega}.$$

□

7.36. Describe the Fourier transform $\mathcal{F}(f)$ of the function

$$f(t) = \mathrm{e}^{-at^2}, \quad t \in \mathbb{R},$$

where $a > 0$.

Solution. Our task is to calculate

$$\mathcal{F}(f)(\omega) = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-at^2}\, \mathrm{e}^{-i\omega t}\, dt.$$

Differentiating (with respect to ω) and then integrating by parts (for $F' = -it\, \mathrm{e}^{-at^2}$, $G = \mathrm{e}^{-i\omega t}$) gives

$$(\mathcal{F}(f)(\omega))' = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} -it\, \mathrm{e}^{-at^2}\, \mathrm{e}^{-i\omega t}\, dt =$$

$$\tfrac{1}{\sqrt{2\pi}} \left(\lim_{t\to\infty} \tfrac{i}{2a}\, \mathrm{e}^{-at^2 - i\omega t} - \lim_{t\to-\infty} \tfrac{i}{2a}\, \mathrm{e}^{-at^2 - i\omega t} - \int\limits_{-\infty}^{\infty} \tfrac{i(-i\omega)}{2a}\, \mathrm{e}^{-at^2}\, \mathrm{e}^{-i\omega t}\, dt \right) =$$

$$\tfrac{1}{\sqrt{2\pi}} \left(\tfrac{i}{2a} \lim_{t\to\infty} \mathrm{e}^{-at^2} - \tfrac{i}{2a} \lim_{t\to-\infty} \mathrm{e}^{-at^2} - \int\limits_{-\infty}^{\infty} \tfrac{\omega}{2a}\, \mathrm{e}^{-at^2}\, \mathrm{e}^{-i\omega t}\, dt \right) =$$

$$-\tfrac{\omega}{2a} \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-at^2}\, \mathrm{e}^{-i\omega t}\, dt \right) = -\tfrac{\omega}{2a}\, \mathcal{F}(f)(\omega).$$

$$= \frac{1}{n} b_n(f').$$

We write $a_n(f)$ for the corresponding coefficient of the function f, and so on.

Thus we can see that the "smoother" a function is, the more rapidly the Fourier coefficients approach zero. 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 the same for $b_n(f)$. In other words, 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 us thus consider a continuous function f 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

$$\begin{aligned} |s_N(x) - f(x)| &= \Big| \sum_{k=N+1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)) \Big| \\ &\le \sum_{k=N+1}^{\infty} (|a_k| + |b_k|). \end{aligned}$$

The right-hand side can further be estimated by the coefficients a'_n and b'_n of the derivative f' (invoking Hölder's inequality for the L_p and L_q norms for infinite series with $p = q = 2$, see 7.15, and Bessel's inequality for general Fourier series, see 7.5.(2))

$$\begin{aligned} |s_N(x) - f(x)| &\le \sum_{k=N+1}^{\infty} \frac{1}{k}(|a'_k| + |b'_k|) \\ &\le \Big(2 \sum_{k=N+1}^{\infty} \frac{1}{k^2}\Big)^{1/2} \Big(\sum_{k=N+1}^{\infty} (|a'_k|^2 + |b'_k|^2)\Big)^{1/2} \\ &\le \sqrt{2}\Big(\int_N^{\infty} \frac{1}{x^2}\, dx\Big)^{1/2} \frac{1}{\sqrt{\pi}} \|f'\|_2 \\ &= \Big(\frac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2\Big) \cdot \frac{1}{\sqrt{N}}. \end{aligned}$$

Thus we have obtained not only a proof of the uniform convergence of our series to the anticipated value, but also a bound for the speed of the convergence:

$$\sup_{x\in\mathbb{R}} |s_N(x) - f(x)| \le \Big(\frac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2\Big) \cdot \frac{1}{\sqrt{N}}.$$

This proves the statement 7.8.(2), supposing the Dirichlet condition 7.8.(1) holds.

7.25. **L_2–convergence.** In the next step of our proof, we will 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 will describe it without further details. Interested readers should be able to fill in the gaps by themselves without any difficulties. First, we will formulate the statement we need in general:

Lemma. *A 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.*

Therefore, let us look for functions $y(\omega) = \mathcal{F}(f)(\omega)$ which satisfy the differential equation

$$(7.33) \qquad y' = -\frac{\omega}{2a}\, y.$$

Writing $y' = dy/d\omega$, we have

$$\tfrac{dy}{d\omega} = -\tfrac{\omega}{2a}\, y, \quad \text{i. e.} \quad \tfrac{1}{y}\, dy = -\tfrac{\omega}{2a}\, d\omega,$$

unless the function y equals zero (apparently, $y \equiv 0$ is a solution of (‖7.33‖)). Integration yields

$$\ln|\,y\,| = -\tfrac{\omega^2}{4a} - \ln|\,C\,|, \quad \text{i. e.} \quad y = \pm\tfrac{1}{C}\,\mathrm{e}^{-\frac{\omega^2}{4a}},$$

where $C \in \mathbb{R}\smallsetminus\{0\}$. Including the zero solution as well, we can express all the solutions of the differential equation (‖7.33‖) as the functions

$$y(\omega) = K\,\mathrm{e}^{-\frac{\omega^2}{4a}}, \quad K \in \mathbb{R}.$$

Let us supplement the determination of the constant K for which we get $\mathcal{F}(f)(\omega)$. Later (in connection with the so-called normal distribution in statistics), we will learn that

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\, dx = \sqrt{\pi},$$

whence it follows that

$$\int_{-\infty}^{\infty} \mathrm{e}^{-at^2}\, dt = \tfrac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\, dx = \tfrac{\sqrt{\pi}}{\sqrt{a}}.$$

Therefore,

$$\mathcal{F}(f)(0) = \tfrac{1}{\sqrt{2\pi}}\,\tfrac{\sqrt{\pi}}{\sqrt{a}} = \tfrac{1}{\sqrt{2a}} \quad \text{and} \quad \mathcal{F}(f)(0) = K\,\mathrm{e}^0 = K.$$

Altogether, we have

$$\mathcal{F}(f)(\omega) = \tfrac{1}{\sqrt{2a}}\,\mathrm{e}^{-\frac{\omega^2}{4a}}.$$

□

7.37. Determine the function f whose Fourier transform is the function

$$\tilde{f}(\omega) = \tfrac{1}{\sqrt{2\pi}}\,\tfrac{\sin\omega}{\omega}, \quad \omega \neq 0.$$

Solution. The inverse Fourier transform gives

$$f(t) = \tfrac{1}{2\pi} \int_{-\infty}^{\infty} \tfrac{\sin\omega}{\omega}\,\mathrm{e}^{i\omega t}\, d\omega =$$

$$\tfrac{1}{2\pi}\left(\int_{-\infty}^{0} \tfrac{\sin\omega}{\omega}\,\mathrm{e}^{i\omega t}\, d\omega + \int_{0}^{\infty} \tfrac{\sin\omega}{\omega}\,\mathrm{e}^{i\omega t}\, d\omega \right).$$

If we substitute $-\omega$ for ω in the integral over the interval $(-\infty, 0]$, we will obtain

$$f(t) = \tfrac{1}{2\pi}\left(\int_{0}^{\infty} \tfrac{\sin\omega}{\omega}\,\mathrm{e}^{-i\omega t}\, d\omega + \int_{0}^{\infty} \tfrac{\sin\omega}{\omega}\,\mathrm{e}^{i\omega t}\, d\omega \right) =$$

$$\tfrac{1}{2\pi} \int_{0}^{\infty} \tfrac{\sin\omega}{\omega}\, \left[\cos(\omega t) - i\sin(\omega t) + \cos(\omega t) + i\sin(\omega t)\right] d\omega =$$

$$\tfrac{1}{\pi} \int_{0}^{\infty} \tfrac{\sin\omega}{\omega}\, \cos(\omega t)\;\, d\omega.$$

Let us mention that the previous expression can be obtained already from the fact that the function $y = \frac{\sin\omega}{\omega}$ with maximal domain is even.

Using the identity

$$\sin x \cdot \cos(xy) = \tfrac{1}{2}\left(\sin\left[x(1+y)\right] + \sin\left[x(1-y)\right]\right), \quad x, y \in \mathbb{R},$$

The idea of the proof can be seen well at the example of approximation of Heaviside's function h on the interval $[-\pi, \pi]$. For every δ satisfying $\pi > \delta > 0$, we define the function f_δ as x/δ for $|x| \le \delta$ and $f_\delta(x) = h(x)$ otherwise. Apparently, all the functions f_δ are continuous since the point of discontinuity was overcome by a convenient linear function on an interval whose size is controlled by δ. It can be calculated very easily that $\|h - f_\delta\|_2 \to 0$ as the function f is bounded in absolute value, and so the contribution of the integration over a decreasing interval has to approach zero.

All discontinuity points of a general function f can be cared 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 of the partial sums of its Fourier series can be bounded with the help of a continuous approach f_ε in this way (all norms in this paragraph are the L_2 norms):

$$\|f - s_N(f)\| \le \|f - f_\varepsilon\| + \|f_\varepsilon - s_N(f_\varepsilon)\| + \|s_N(f_\varepsilon) - s_N(f)\|$$

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 equally tightly. It is good to notice that the third one is the size of the partial sum of the Fourier series for $f - f_\varepsilon$. Thus we have

$$\|f - f_\varepsilon - s_N(f - f_\varepsilon)\| \le \|f - f_\varepsilon\|.$$

Therefore, (thanks to the triangle inequality)

$$\|s_N(f - f_\varepsilon)\| \le 2\|f - f_\varepsilon\| \le 2\varepsilon.$$

Altogether, we have bounded the whole distance for sufficiently close continuous functions and sufficiently large numbers N by the number 4ε. This verifies the L_2 convergence we wanted to prove.

7.26. Dirichlet kernel. Finally, we get to the proof of the first statement of theorem 7.8. It follows straight from the definition of the Fourier series $F(t)$ for a function $f(t)$, using its expression with the complex exponential in 7.7, that the partial sums $s_N(t)$ can be written as

$$s_N(t) = \frac{1}{T} \sum_{k=-N}^{N} \int_{-T/2}^{T/2} f(x)\,\mathrm{e}^{-i\omega kx}\,\mathrm{e}^{i\omega kt}\, dx,$$

where T is the period we are working with and $\omega = 2\pi/T$. This expression can be rewritten as

$$s_N(t) = \int_{-T/2}^{T/2} K_N(t-x) f(x)\, dx,$$

and the function

$$K_N(y) = \frac{1}{T} \sum_{k=-N}^{N} \mathrm{e}^{i\omega ky}$$

is called *Dirichlet kernel*. Let us notice that the sum is a piece of a geometric series with common ratio $\mathrm{e}^{i\omega y}$. Thus it can be expressed explicitly for all $y \neq 0$ in the following way (both the numerator which, besides others, follows from the sum formulae (for the sine function), we get

$$f(t) = \tfrac{1}{2\pi}\left(\int\limits_0^\infty \tfrac{\sin[\omega(1+t)]}{\omega}\,d\omega + \int\limits_0^\infty \tfrac{\sin[\omega(1-t)]}{\omega}\,d\omega\right).$$

The substitutions $u = \omega\,(1+t)$, $v = \omega\,(1-t)$ then give

$$f(t) = \tfrac{1}{2\pi}\left(\int\limits_0^\infty \tfrac{\sin u}{u}\,du - \int\limits_0^\infty \tfrac{\sin v}{v}\,dv\right) = 0, \quad t > 1;$$

$$f(t) = \tfrac{1}{2\pi}\left(\int\limits_0^\infty \tfrac{\sin u}{u}\,du + \int\limits_0^\infty \tfrac{\sin v}{v}\,dv\right) = \tfrac{1}{\pi}\int\limits_0^\infty \tfrac{\sin u}{u}\,du, \quad t \in (-1, 1);$$

$$f(t) = \tfrac{1}{2\pi}\left(-\int\limits_0^\infty \tfrac{\sin u}{u}\,du + \int\limits_0^\infty \tfrac{\sin v}{v}\,dv\right) = 0, \quad t < -1.$$

Thus we have proved that the function f is zero for $|\,t\,| > 1$ and constant (necessarily non-zero) for $|\,t\,| < 1$. (All the way, we assume that the inverse Fourier transform exists.)

Let us determine the function value $f(0)$. The function

$$g(t) = 1, \quad |\,t\,| < 1; \qquad g(t) = 0, \quad |\,t\,| > 1$$

satisfies

$$\mathcal{F}(g)(\omega) = \tfrac{1}{\sqrt{2\pi}}\int\limits_{-1}^{1} \mathrm{e}^{-i\omega t}\,dt = \tfrac{2}{\sqrt{2\pi}}\int\limits_0^1 \cos(\omega t)\,dt = \tfrac{2}{\sqrt{2\pi}}\tfrac{\sin\omega}{\omega}.$$

Hence it follows that $f(0) = g(0)/2 = 1/2$. Let us emphasize enumeration of the integral

$$\int\limits_0^\infty \tfrac{\sin u}{u}\,du = \tfrac{\pi}{2},$$

which we have obtained as well. □

7.38. Solve the integral equation

$$\int\limits_0^\infty f(x)\,\sin(xt)\,dt = \mathrm{e}^{-x}, \quad x > 0$$

for an unknown function f.

Solution. If we multiply both sides of the equation by the number $\sqrt{2/\pi}$, we obtain just the sine Fourier transform on the left-hand side. Therefore, it suffices to apply the inverse transform to the equation. Thus we get

$$f(t) = \tfrac{2}{\pi}\int\limits_0^\infty \mathrm{e}^{-x}\,\sin(xt)\,dx, \quad t > 0.$$

Integrating by parts twice, we can obtain

$$\int \mathrm{e}^{-x}\,\sin(xt)\,dx = \tfrac{\mathrm{e}^{-x}}{1+t^2}\left[-\sin(xt) - t\cos(xt)\right] + C,$$

hence

$$\int\limits_0^\infty \mathrm{e}^{-x}\,\sin(xt)\,dx =$$

$$\lim_{x\to\infty}\left(\tfrac{\mathrm{e}^{-x}}{1+t^2}\left[-\sin(xt) - t\cos(xt)\right]\right) - \tfrac{\mathrm{e}^0}{1+t^2}(-t) = \tfrac{t}{1+t^2}.$$

Therefore, the function

$$f(t) = \tfrac{2}{\pi}\,\tfrac{t}{1+t^2}, \quad t > 0.$$

is the solution of the equation. □

and the denominator are multiplied by $-\,\mathrm{e}^{-i\omega y/2}$ to be able to substitute the real-valued function sin):

$$\begin{aligned} K_N(y) &= \frac{1}{T}\frac{\mathrm{e}^{-iN\omega y} - \mathrm{e}^{i(N+1)\omega y}}{1 - \mathrm{e}^{i\omega y}} \\ &= \frac{1}{T}\frac{-\,\mathrm{e}^{-i(N+1/2)\omega y} + \mathrm{e}^{i(N+1/2)\omega y}}{\mathrm{e}^{i\omega y/2} - \mathrm{e}^{-i\omega y/2}} \\ &= \frac{1}{T}\frac{\sin((N+1/2)\omega y)}{\sin(\omega y/2)}. \end{aligned}$$

At the point $y = 0$, of course, we see that $K_N(0) = \frac{1}{T}(2N+1)$.

It is apparent from the last expression that $K_N(y)$ is an even function, and using l'Hospital's rule, we can calculate quickly that it is continuous everywhere. Since all the partial sums of the series for the constant function $f(x) = 1$ also equal 1, we get from the definition of the Dirichlet kernel that

$$\int_{-T/2}^{T/2} K_N(x)dx = 1.$$

In the case of periodic functions, the integrals over intervals whose length equals the period are independent of the choice of the marginal 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 will focus on 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 $F(x)$ converges to the value $f(x)$ at the point x. We get

$$s_N(x) - f(x) = \int_{-T/2}^{T/2} (f(x+y) - f(x))K_N(y)\,dy.$$

The integrated expression can be rewritten into a form which reminds Fourier coefficients of convenient functions:

$$\begin{aligned} &\frac{f(x+y) - f(x)}{\sin(\omega y/2)}\sin((N+1/2)\omega y) = \\ &\quad = \varphi_x(y)(\cos(\omega y/2)\sin(N\omega y) + \sin(\omega y/2)\cos(N\omega y)), \end{aligned}$$

where we used the denotation

$$\varphi_x(y) = \frac{f(x+y) - f(x)}{\sin(\omega y/2)}$$

for $y \neq 0$, while $\varphi_x(0) = f'(x)$. Let us notice that we needed the differentiability and continuity of f at the point x in this step of the proof.

Now, we can truly perceive the difference $s_N(x) - f(x)$ as the sum of Fourier coefficients $b_N(\psi_1)$ and $a_N(\psi_2)$ where

$$\psi_1(y) = \frac{T}{2}\varphi_x(y)\cos(\omega y/2), \;\; \psi_2(y) = \frac{T}{2}\varphi_x(y) sin(\omega y/2).$$

However, this means that as N increases, this expression $b_N(\psi_1) + a_N(\psi_2)$ necessarily converges to zero (see 7.5.(2)).

For the end, we will look at the convergence in the case when the function f or its derivative has a discontinuity point at $x = 0$. Since the function belongs to $\mathcal{S}^1$, it is already continuous and differentiable on a neighborhood of the point $x = 0$ (outside the

7.39. 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 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

$$\begin{aligned} I &= c\varepsilon_0 \frac{1}{\tau}\int_0^\tau E_y^2 \,\mathrm{d}t = c\varepsilon_0 \frac{1}{\tau}\int_0^\tau A^2\cos^2(\omega t - kx)\,\mathrm{d}t \\ &= c\varepsilon_0 A^2 \frac{1}{\tau}\int_0^\tau \frac{1+\cos(2(\omega t - kx))}{2}\,\mathrm{d}t \\ &= \frac{1}{2}c\varepsilon_0 A^2 \frac{1}{\tau}\big[t + \frac{\sin(2(\omega t - kx))}{2\omega}\big]_0^\tau \\ &= \frac{1}{2}c\varepsilon_0 A^2 \frac{1}{\tau}\big(\tau + \frac{\sin(2(\omega\tau - kx)) - \sin(2(-kx))}{2\omega}\big) \\ &= \frac{1}{2}c\varepsilon_0 A^2 \big(1 + \frac{\sin(2(\omega\tau - kx)) - \sin(2(-kx))}{2\omega\tau}\big) \doteq \frac{1}{2}c\varepsilon_0 A^2 \end{aligned}$$

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):

point itself). Let us split the f into its even part f_1 and its odd part f_2, i. e.,

$$f(x) = \frac{1}{2}(f(x) + f(-x)) + \frac{1}{2}(f(x) - f(-x)).$$

The value $f_1(0)$ at the point $x = 0$ is then defined as

$$\frac{1}{2}(\lim_{y\to 0_+} f(y) + \lim_{y\to 0_-} f(y)).$$

Then we can easily verify that the even part $f_1(x)$ is continuous and differentiable at the point $x = 0$ (thanks to existence of the one-sided limits), and so on the entire neighborhood of this point. We are surely not surprised by the fact that the odd part satisfies $f_2(0) = 0$, and so does the Fourier series which 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 our function f, the identity

$$F(0) = F_1(0) + F_2(0) = \frac{1}{2}\big(\lim_{y\to 0_+} f(y) + \lim_{y\to 0_-} f(y)\big) + 0,$$

which we wanted to prove.

In the case of discontinuity at a general point, we can proceed similarly, and the whole proof has come to an end (so has the proof of the statements (2) and (3) of theorem 7.8 where we required that the Dirichlet condition be true).

3. Integral operators

7.27. Integral operators. In the case of finite-dimensional vector spaces, we can perceive 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. Then we 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 mapped vectors to scalars (the so-called linear forms). It was defined as the sum of products of coordinates x_i of vectors with fixed values $\alpha_i = \alpha(e_i)$ at the generators e_i, i. e. by one-row matrices:

$$(x_1, \dots, x_n)^T \mapsto (\alpha_1, \dots, \alpha_n)\cdot(x_1, \dots, x_n)^T.$$

More complicated mappings, with values lying in the same space, were then given similarly by square matrices. We can approach linear operations on spaces of functions in an analogous way.

For the sake of simplicity, we will work with the real vector space $\mathcal{S}$ of all piecewise continuous real-valued functions having a compact support and defined on the whole $\mathbb{R}$ or on an interval $I = [a, b]$. Linear mappings $\mathcal{S} \to \mathbb{R}$ will be 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.,

$$L(f) = f(x_0).$$

Or, we can have the functional given by integration and a fixed function $g(x)$, i. e.,

$$L(f) = \int_a^b f(x)g(x)\,dx.$$

$$\psi(\xi,\eta)=A\iint_{\mathbb{R}^2} s(x,y)e^{-ik(\xi x+\eta y)}\,\mathrm{d}x\,\mathrm{d}y$$

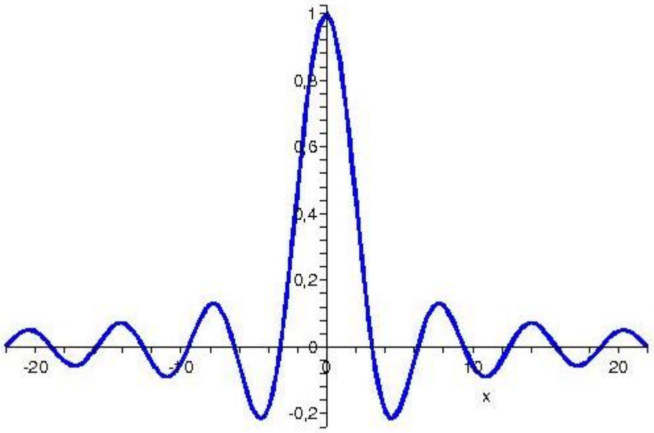
.

$$\psi(\xi,\eta)=A\int_{-p/2}^{p/2}\int_{-q/2}^{q/2} e^{-ik(\xi x+\eta y)}\,\mathrm{d}y\,\mathrm{d}x$$

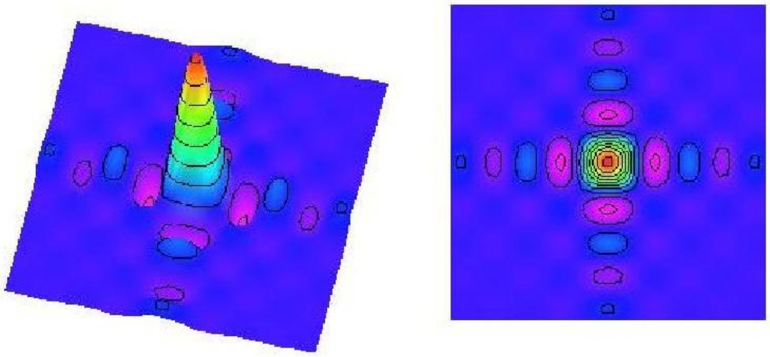
$$\psi(\xi,\eta)=A\int_{-p/2}^{p/2} e^{-ik\xi x}\,\mathrm{d}x\int_{-q/2}^{q/2} e^{-ik\eta y}\,\mathrm{d}y=A\left[\frac{e^{-ik\xi x}}{-ik\xi}\right]_{-p/2}^{p/2}\left[\frac{e^{-ik\eta y}}{-ik\eta}\right]_{-q/2}^{q/2}$$

$$=A\frac{2\sin(k\xi p/2)}{k\xi}\frac{2\sin(k\eta q/2)}{k\eta}=Apq\frac{\sin(k\xi p/2)}{k\xi p/2}\frac{\sin(k\eta q/2)}{k\eta q/2}$$

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:

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.

We can get a good image from the choice

$$g(x)=\begin{cases}0 & \text{if } |x|\geq\varepsilon\\ \frac{1}{2\varepsilon} & \text{if } |x|<\varepsilon\,.\end{cases}$$

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–neighborhood of the origin. Similarly, we can work with the function

$$g(x)=\begin{cases}0 & \text{if } |x|\geq\varepsilon\\ \mathrm{e}^{\frac{1}{x^2-\varepsilon^2}+\frac{1}{\varepsilon^2}} & \text{if } |x|<\varepsilon\end{cases}$$

with which worked in the paragraph 6.6. This function is smooth on the whole $\mathbb{R}$ with a 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. Surely, the integral of g over the whole $\mathbb{R}$ is finite, yet 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.

There is another quite common instance, the so-called Gaussian function

$$g(x)=\frac{1}{\pi}\,\mathrm{e}^{-x^2},$$

which also has a unit integral over the whole $\mathbb{R}$ (we will 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.

We could observe another example with a unit integral over the whole a while ago when we were discussing the Dirichlet kernels $g(x)=K_N(x)$ for Fourier series.

7.28. Function convolution. Integral functionals from the previous paragraph can easily be modified to obtain a "steamed averaging" of the values of a given function f near a given point $y\in\mathbb{R}$:

$$L_y(f)=\int_{-\infty}^{\infty} f(x)g(y-x)\,dx$$

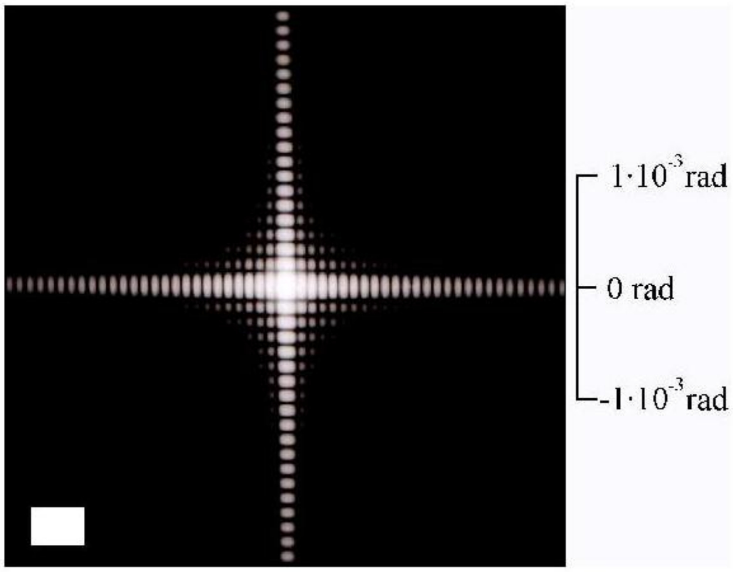
Convolution of functions of a real variable

The free parameter y in the definition of the functional $L_y(f)$ can be perceived as a new independent variable, and our operation L_y actually maps functions to functions again, $f\mapsto\tilde{f}$:

$$\tilde{f}(y)=L_y(f)=\int_{-\infty}^{\infty} f(x)g(y-x)\,dx.$$

This operation is called the *convolution of functions* f and g, denoted $f*g$.

The convolution is mostly defined for real or complex functions on $\mathbb{R}$ with a compact support.



Since $\lim_{x\to0}\frac{\sin x}{x}=1$, the intensity at the middle of the image is directly proportional to $I_0=A^2p^2q^2$. 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.

7.40. Find the solution to the so-called equation of heat conduction (equation of diffusion)

$$u_t(x,t)=a^2\,u_{xx}(x,t),\qquad x\in\mathbb{R},\ t>0$$

satisfying the initial condition $\lim\limits_{t\to0+}u(x,t)=f(x)$.

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, $u_{xx}=\frac{\partial^2u}{\partial x^2}$ 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

$$\mathcal{F}\left(u_t\right)(\omega,t)=\tfrac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}u_t(x,t)\,\mathrm{e}^{-i\omega x}\,dx=$$

$$\left(\tfrac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}u(x,t)\,\mathrm{e}^{-i\omega x}\,dx\right)',$$

where differentiated with respect to t, i. e.,

$$\mathcal{F}\left(u_t\right)(\omega,t)=\left(\mathcal{F}\left(u\right)(\omega,t)\right)'=\left(\mathcal{F}\left(u\right)\right)_t(\omega,t).$$

At the same time, we know that

By the transformation $t=z-x$, we can easily calculate that

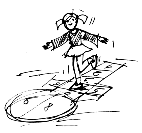
$$\begin{aligned}(f*g)(z)&=\int_{-\infty}^{\infty}f(x)g(z-x)\,dx\\&=-\int_{\infty}^{-\infty}f(z-t)g(t)\,dt=(g*f)(z).\end{aligned}$$

Thus the convolution, considered a binary operation

$$*:\mathcal{S}_c\times\mathcal{S}_c\to\mathcal{S}_c$$

of pairs of functions having compact supports, is commutative. Similarly, convolutions can be considered with integration over a finite interval; we only have to guarantee that the functions participating in them be well-defined. Especially, this can be done for periodic functions with integrating over an interval whose length equals the period.

The convolution is an extraordinarily useful tool for modeling the way in which we observe the data of an experiment or the influence of a medium through which information is transferred (for instance, an analog audio or video signal affected by noise, and so on) The input value f is the transferred information and 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.



7.29. Gibbs phenomenon. Actually, we have already seen a useful case of convolution. In paragraph 7.26, we interpreted the partial sum of the Fourier series for a function f as a convolution with Dirichlet kernel $K_N(y)=\sum_{-T/2}^{T/2}\mathrm{e}^{i\omega ky}$.

This interpretation allows us to explain the so-called Gibbs phenomenon mentioned in paragraph 7.9.

7.30. Fourier transform. The convolution is one of many examples of a general integral 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 domain of such functionals must be chosen in view of the properties of the particular kernels so that the used integral would exist at all.

The theory of integral operators with kernels and equations they contain is very useful and interesting at the same time. However, we do not have enough space for it here. We will focus only on an extraordinarily important case, the so-called *Fourier transform $\mathcal{F}$*, which has deep connections with Fourier series.

Let us remind that a function $f(t)$, given by its converging Fourier series, equals

$$f(t)=\sum_{n=-\infty}^{\infty}c_n\,\mathrm{e}^{i\omega_n t},$$

where the numbers c_n are complex Fourier coefficients, $\omega_n=n2\pi/T$ with period T, see paragraph 7.7.

Having fixed T, the expression $\Delta\omega=2\pi/T$ describes just 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

$$\mathcal{F}\left(a^2 u_{xx}\right)(\omega, t) = a^2\,\mathcal{F}\left(u_{xx}\right)(\omega, t) = -a^2\omega^2\,\mathcal{F}\left(u\right)(\omega, t).$$

Denoting $y(\omega, t) = \mathcal{F}\left(u\right)(\omega, t)$, 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

$$y(\omega, t) = K(\omega)\, \mathrm{e}^{-a^2\omega^2 t}, \quad K(\omega) \in \mathbb{R}.$$

It remains to determine $K(\omega)$. The transformation of the initial condition gives

$$\mathcal{F}\left(f\right)(\omega) = \lim_{t\to 0+} \mathcal{F}\left(u\right)(\omega, t) = \lim_{t\to 0+} y(\omega, t) = K(\omega)\,\mathrm{e}^0 = K(\omega),$$

hence

$$y(\omega, t) = \mathcal{F}\left(f\right)(\omega)\, \mathrm{e}^{-a^2\omega^2 t}, \quad K(\omega) \in \mathbb{R}.$$

Now, using the inverse Fourier transform, we can return to the original differential equation with solution

$$u(x, t) = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} y(\omega, t)\, \mathrm{e}^{i\omega x}\, d\omega =$$

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathcal{F}\left(f\right)(\omega)\, \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{i\omega x}\, d\omega =$$

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s)\, \mathrm{e}^{-i\omega s}\, ds \right) \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{i\omega x}\, d\omega =$$

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s) \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{-i\omega(s-x)}\, d\omega \right) ds.$$

Computing the Fourier transform $\mathcal{F}(f)$ of the function $f(t) = \mathrm{e}^{-at^2}$ for $a > 0$, we have obtained (while relabeling the variables)

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-cp^2}\, \mathrm{e}^{-irp}\, dp = \tfrac{1}{\sqrt{2c}}\, \mathrm{e}^{-\frac{r^2}{4c}}, \quad c > 0.$$

According to this formula (consider $c = a^2 t > 0$, $p = \omega$, $r = s - x$), we have

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{-i\omega(s-x)}\, d\omega = \tfrac{1}{\sqrt{2a^2 t}}\, \mathrm{e}^{-\frac{(s-x)^2}{4a^2 t}}.$$

Therefore,

$$u(x, t) = \tfrac{1}{2a\sqrt{\pi t}} \int\limits_{-\infty}^{\infty} f(s)\, \mathrm{e}^{-\frac{(x-s)^2}{4a^2 t}}\, ds.$$

□

7.41. Determine the Laplace transform $\mathcal{L}(f)(s)$ of the function

(a) $f(t) = \mathrm{e}^{at}$;
(b) $f(t) = c_1\,\mathrm{e}^{a_1 t} + c_2\,\mathrm{e}^{a_2 t}$;
(c) $f(t) = \cos(bt)$;
(d) $f(t) = \sin(bt)$;
(e) $f(t) = \cosh\,(bt)$;
(f) $f(t) = \sinh\,(bt)$,

where the values $b \in \mathbb{R}$ and $c_1, c_2 \in \mathbb{C}$ are arbitrary and 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 it is also greater than b in the problems (e) and (f).

Solution. The case (a). It follows directly from the definition of the Laplace transform that

calculating the coefficients of the Fourier series. The coefficient $1/T$ in the formula

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-i\omega_n t}\, dt$$

then equals $\Delta\omega/2\pi$, so the series for $f(t)$ can be rewritten as

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi}\left(\Delta\omega \int_{-T/2}^{T/2} f(x)\,\mathrm{e}^{-i\omega_n x}\, dx\, \mathrm{e}^{i\omega_n t}\right).$$

Now, let us 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 $f(t)$ actually describes the summands of the Riemann sums for the improper integral

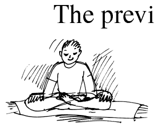
$$\frac{1}{2\pi}\int_{-\infty}^{\infty} g(\omega)\,\mathrm{e}^{i\omega t}\, d\omega,$$

where $g(\omega)$ is a function which takes, at the points ω_n, the values

$$g(\omega_n) = \int_{-T/2}^{T/2} f(x)\,\mathrm{e}^{-i\omega_n x}\, dx.$$

We are working with piecewise continuous functions with a compact support, thus our function f is integrable in absolute value over the whole $\mathbb{R}$. Letting $T \to \infty$, the norm $\Delta\omega$ of our subintervals in the Riemann sum gets finer. At the same time, in the last expression, we obtain the integral

$$g(\omega) = \int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{-i\omega x}\, dx.$$



The previous reasonings show that there is quite 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 a compact support, we define

$$\mathcal{F}(f)(\omega) = \tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-i\omega t}\, dt.$$

This function $\tilde{f}$ is called the Fourier transform of the function f. The previous reasonings also show that we will have

$$f(t) = \mathcal{F}^{-1}(\tilde{f})(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \tilde{f}(\omega)\,\mathrm{e}^{i\omega t}\, d\omega.$$

This says that the *Fourier transform* $\mathcal{F}$ just defined has an inverse operation $\mathcal{F}^{-1}$, which is called *inverse Fourier transform*.

Let us notice that both the Fourier transform and its inverse are integral operators with almost identical kernels

$$k(\omega, t) = \mathrm{e}^{\pm i\omega t}.$$

Of course, there transforms are meaningful for much greater domains. Interested readers are referenced to specialized literature.

$$\mathcal{L}(f)(s) = \int\limits_0^\infty \mathrm{e}^{at}\,\mathrm{e}^{-st}\,dt = \int\limits_0^\infty \mathrm{e}^{-(s-a)t}\,dt =$$
$$\lim_{t\to\infty}\left(\frac{\mathrm{e}^{-(s-a)t}}{-(s-a)}\right) - \frac{\mathrm{e}^0}{-(s-a)} = \frac{1}{s-a}.$$

The case (b). Using the result of the above case and the linearity of improper integrals, we obtain

$$\mathcal{L}(f)(s) = c_1 \int\limits_0^\infty \mathrm{e}^{a_1 t}\,\mathrm{e}^{-st}\,dt + c_2 \int\limits_0^\infty \mathrm{e}^{a_2 t}\,\mathrm{e}^{-st}\,dt = \frac{c_1}{s-a_1} + \frac{c_2}{s-a_2}.$$

The case (c). Since

$$\cos(bt) = \tfrac{1}{2}\left(\mathrm{e}^{ibt} + \mathrm{e}^{-ibt}\right),$$

the choice $c_1 = 1/2 = c_2$, $a_1 = ib$, $a_2 = -ib$ in the previous variant gives

$$\mathcal{L}(f)(s) = \int\limits_0^\infty \left(\tfrac{1}{2}\mathrm{e}^{ibt} + \tfrac{1}{2}\mathrm{e}^{-ibt}\right)\mathrm{e}^{-st}\,dt = \frac{1}{2(s-ib)} + \frac{1}{2(s+ib)} = \frac{s}{s^2+b^2}.$$

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) $\mathcal{L}(f)(s) = \frac{b}{s^2-b^2}$.

□

7.42. Using the relation

$$\mathcal{L}\left(f'\right)(s) = s\,\mathcal{L}(f)(s) - \lim_{t\to 0+} f(t), \tag{7.34}$$

derive the Laplace transforms of the functions $y = \cos t$ and $y = \sin t$.

Solution. First, let us realize that from (‖7.34‖), it follows that

$$\mathcal{L}\left(f''\right)(s) = s\,\mathcal{L}\left(f'\right)(s) - \lim_{t\to 0+} f'(t) =$$
$$s\left(s\mathcal{L}(f)(s) - \lim_{t\to 0+} f(t)\right) - \lim_{t\to 0+} f'(t) =$$
$$s^2\mathcal{L}(f)(s) - s\lim_{t\to 0+} f(t) - \lim_{t\to 0+} f'(t).$$

Therefore,

$$-\mathcal{L}(\sin t)(s) = \mathcal{L}(-\sin t)(s) = \mathcal{L}\left((\sin t)''\right)(s) =$$
$$s^2\mathcal{L}(\sin t)(s) - s\lim_{t\to 0+}\sin t - \lim_{t\to 0+}\cos t = s^2\mathcal{L}(\sin t)(s) - 1,$$

whence we get

$$-\mathcal{L}(\sin t)(s) = s^2\mathcal{L}(\sin t)(s) - 1, \quad \text{i. e.} \quad \mathcal{L}(\sin t)(s) = \frac{1}{s^2+1}.$$

Now, invoking (‖7.34‖), we can easily determine

$$\mathcal{L}(\cos t)(s) = \mathcal{L}\left((\sin t)'\right)(s) = s\,\frac{1}{s^2+1} - \lim_{t\to 0+}\sin t = \frac{s}{s^2+1}.$$

□

7.43. For $s > -1$, calculate the Laplace transform $\mathcal{L}(g)(s)$ of the function

$$g(t) = t\,\mathrm{e}^{-t}.$$

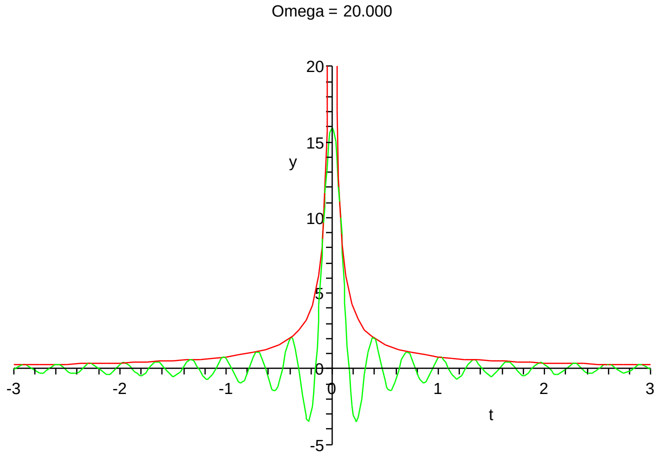
Further, for $s > 1$, calculate the Laplace transform $\mathcal{L}(h)(s)$ of the function

7.31. Simple properties. The Fourier transform changes the local and global behavior of functions in an interesting way. Let us begin with a simple example in which we find 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 $|\omega| \le \Omega$. The inverse transform $\mathcal{F}^{-1}$ gives

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\Omega}^{\Omega} \mathrm{e}^{i\omega t}\,d\omega = \frac{1}{\sqrt{2\pi}}\left[\frac{1}{it}\,\mathrm{e}^{i\omega t}\right]_{-\Omega}^{\Omega} \\ &= \frac{2}{\sqrt{2\pi}\,t}\,\frac{1}{2i}(\mathrm{e}^{i\Omega t} - \mathrm{e}^{-i\Omega t}) \\ &= \frac{2\Omega}{\sqrt{2\pi}}\,\frac{\sin(\Omega t)}{\Omega t}. \end{aligned}$$

Thus, except for a multiplicative constant and the scaling of the input variable, it is the very important function $\mathrm{sinc}(x) = \frac{\sin x}{x}$.

Straight calculation of the limit at zero (l'Hospital's rule) gives $f(0) = 2\Omega(2\pi)^{-1/2}$, the closest zero points are at $t = \pm\pi/\Omega$ and the function drops to zero quite rapidly outside the origin $x = 0$. This function is caught in the picture 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.



We can see 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.

Further, we will derive the Fourier transform of the derivative $f'(t)$ for a function f. We keep supposing that f has a compact support, i. e., especially both $\mathcal{F}(f')$ and $\mathcal{F}(f)$ really exist. Let us use integration by parts:

$$\begin{aligned} \mathcal{F}(f')(\omega) &= \frac{1}{\sqrt{2\pi}}\int_{\infty}^{\infty} f'(t)\,\mathrm{e}^{-i\omega t}\,dt \\ &= \frac{1}{\sqrt{2\pi}}\left[\mathrm{e}\,-i\omega t\,f(t)\right]_{-\infty}^{\infty} + \frac{i\omega}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-i\omega t}\,dt \\ &= i\omega\mathcal{F}(f)(\omega). \end{aligned}$$

Thus we can see that Fourier transform converts the (limit) operation of differentiation to the (algebraic) operation of a simple multiplication by the variable. Of course, this formula can be iterated, obtaining

$$\mathcal{F}(f'')(\omega) = -\omega^2\mathcal{F}(f), \ldots, \mathcal{F}(f^{(n)}) = i^n\,\omega^n\mathcal{F}(f).$$

$$h(t) = t \sinh t.$$

Solution. Integrating by parts, we obtain

$$\mathcal{L}(g)(s) = \int\limits_0^\infty t\,\mathrm{e}^{-t}\,\mathrm{e}^{-st}\,dt = \int\limits_0^\infty t\,\mathrm{e}^{-(s+1)t}\,dt = \lim_{t\to\infty}\left(\tfrac{t\,\mathrm{e}^{-(s+1)t}}{-(s+1)}\right) - 0 -$$

$$\int\limits_0^\infty \tfrac{\mathrm{e}^{-(s+1)t}}{-(s+1)}\,dt = -\left(\lim_{t\to\infty}\tfrac{\mathrm{e}^{-(s+1)t}}{(s+1)^2} - \tfrac{\mathrm{e}^0}{(s+1)^2}\right) = \tfrac{1}{(s+1)^2}.$$

Differentiating the Laplace transform of a general function $-f$ (i. e., an improper integral) with respect to the parameter s gives

$$\left(\int\limits_0^\infty -f(t)\,\mathrm{e}^{-st}\,dt\right)' = \int\limits_0^\infty -f(t)\,\left(\mathrm{e}^{-st}\right)'\,dt = \int\limits_0^\infty t\,f(t)\,\mathrm{e}^{-st}\,dt.$$

This means that the derivative of the Laplace transform $\mathcal{L}(-f)(s)$ is the Laplace transform of the function $tf(t)$. The Laplace transform of the function $y = \sinh t$ has already been determined as the function $y = \frac{1}{s^2-1}$. Therefore,

$$\mathcal{L}(h)(s) = \left(-\tfrac{1}{s^2-1}\right)' = \tfrac{2s}{(s^2-1)^2}.$$

Let us notice that we could also have determined $\mathcal{L}(g)(s)$ this way. □

The basic Laplace transforms are enumerated in the following table:

$y(t)$	$\mathcal{L}(y)(s)$
e^{at}	$\frac{1}{s-a}$
te^{at}	$\frac{1}{(s-a)^2}$
$t^n\,e^{at}$	$\frac{n!}{(s-a)^{n+1}}$
$\sin\omega t$	$\frac{\omega}{s^2+\omega^2}$
$\cos\omega t$	$\frac{s}{s^2+\omega^2}$
$e^{at}\sin\omega t$	$\frac{\omega}{(s-a)^2+\omega^2}$
$e^{at}(\cos\omega t + \frac{a}{\omega}\sin\omega t)$	$\frac{s}{(s-a)^2+\omega^2}$
$t\sin\omega t$	$\frac{2\omega s}{(s^2+\omega^2)^2}$
$\sin\omega t - \omega t\cos\omega t$	$\frac{2\omega^3}{(s^2+\omega^2)^2}$

7.44. Prove the 4th and 5th rows of the table using Euler's formula $e^{i\omega t} = \cos\omega t + i\sin\omega t$.

Solution.

$$\begin{aligned}
&\mathcal{L}(\cos\omega t)(s) + i\mathcal{L}(\sin\omega t)(s) = \mathcal{L}(e^{i\omega t})(s)\\
&= \int_0^\infty e^{i\omega t}e^{-st}\,\mathrm{d}t = \int_0^\infty e^{(i\omega - s)t}\,\mathrm{d}t\\
&= -\frac{1}{s-i\omega}\left[e^{(i\omega-s)t}\right]_0^\infty\\
&= -\frac{1}{s-i\omega}(\lim_{t\to\infty}\frac{e^{i\omega t}}{e^{st}} - 1) = \frac{1}{s-i\omega} = \frac{s+i\omega}{(s-i\omega)(s+i\omega)}\\
&= \frac{s}{s^2+\omega^2} + i\frac{\omega}{s^2+\omega^2}
\end{aligned}$$

□

7.45. Let $\mathcal{L}(y)(s)$ denote the Laplace transform of a function $y(t)$. Using integration by parts, prove that

Solution.

$$\mathcal{L}(y')(s) = s\mathcal{L}(y)(s) - y(0) \qquad (7.35)$$

7.32. The relation to convolutions. There is another extremely important property, the relation between convolutions and Fourier transforms. Let us calculate the transform of the convolution $h = f * g$, where, as usual, we assume that the functions have compact supports. We will switch the order of integration, which is a step whose correctness will be verified later in differential and integral calculus, see **??**. In the next little step, we will introduce the substitution $t - x = u$.

$$\begin{aligned}
\mathcal{F}(h)(\omega) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(x)g(t-x)\,dx\right)\mathrm{e}^{-i\omega t}\,dt\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{\infty} g(t-x)\,\mathrm{e}^{-i\omega t}\,dt\right)dx\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{\infty} g(u)\,\mathrm{e}^{-i\omega(u+x)}\,du\right)dx\\
&= \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{-i\omega x}\,dx\right)\cdot\left(\int_{-\infty}^{\infty} g(u)\,\mathrm{e}^{-i\omega u}\,du\right)\\
&= \sqrt{2\pi}\mathcal{F}(f)\cdot\mathcal{F}(g)
\end{aligned}$$

A similar calculation shows the reverse statement, i. e., the fact that Fourier transform of a product is the convolution of the transforms, up to a constant.

$$\mathcal{F}(f\cdot g) = \frac{1}{\sqrt{2\pi}}\mathcal{F}(f) * \mathcal{F}(g).$$

As we mentioned above, the convolution $f * g$ very often models the process of our observation of some quantity f. Using the Fourier transform and its inverse, we can now easily recognize the original values of this quantity 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. We talk about *deconvolution.*

7.33. Dirac delta–function. Now, let us 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 coveted 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 goes to ∞, the left-hand expression transforms to $\mathcal{F}^{-1}(\mathcal{F}(g))(z) = g(z)$, while on the right-hand side, we get

$$g(z) = \int_{-\infty}^{\infty} g(t)\delta(z-t)\,dt.$$

The wanted $\delta(t)$ thus looks as a "function" which takes zero everywhere except the single point $t = 0$ where it "takes such an infinite value" that integrating the product of $\delta(t)$ and any integrable function g gives just the value of g at the point $t = 0$. Of course, it is not a function in the common sense, but it is an object used quite often. It is called the *Dirac function* δ and it can be described correctly as an instance of the so-called distribution. Since we do not have enough space and time, we will not pay further attention to distributions. We only mention that the Dirac δ can be imagined as a unit impulse at a single point. Its Fourier transform is the constant function $\mathcal{F}(\delta)(\omega) = \frac{1}{\sqrt{2\pi}}$.

$$\mathcal{L}(y'')(s) = s^2\,\mathcal{L}(y) - sy(0) - y'(0)$$

and, by induction:

$$\mathcal{L}(y^{(n)})(s) = s^n\,\mathcal{L}(y)(s) - \sum_{i=1}^{n} s^{n-i}\,y^{(i-1)}(0).$$

□

7.46. 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 previous example ||8.149||:

$$s^2\,\mathcal{L}(y)(s) + 4\mathcal{L}(y)(s) = \mathcal{L}(\sin 2t)(s)$$

At the same time,

$$\mathcal{L}(\sin 2t)(s) = \frac{2}{s^2+4},$$

i. e.,

$$\mathcal{L}(y)(s) = \frac{2}{(s^2+4)^2}.$$

The inverse transform gives

$$y(t) = \tfrac{1}{8}\sin 2t - \tfrac{1}{4}t\cos 2t.$$

□

7.47. Find a function $y(t)$ satisfying the differential equation

$$y''(t) + 6y'(t) + 9y(t) = 50\sin t$$

and the initial conditions $y(0) = 1$ and $y'(0) = 4$.

Solution. The Laplace transform gives

$$s^2\mathcal{L}(y)(s) - s - 4 + 6(s\mathcal{L}(y)(s) - 1) + 9\mathcal{L}(y)(s) = 50\mathcal{L}(\sin t)(s),$$

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 into partial fractions, we get

$$\frac{50}{(s^2+1)(s+3)^2} = \frac{As+B}{s^2+1} + \frac{C}{s+3} + \frac{D}{(s+3)^2},$$

so

$$50 = (As+B)(s+3)^2 + C(s^2+1)(s+3) + D(s^2+1).$$

Substituting $s = -3$, we obtain

$$50 = 10D \quad \text{so} \quad D = 5,$$

and confronting the coefficients at s^3

$$0 = A + C, \quad \text{so} \quad A = -C.$$

Confronting the coefficients at s then yields

$$0 = 9A + 6B + C = 8A + 6B, \quad \text{so} \quad B = \frac{4}{3}C.$$

Confronting the absolute terms, we get

$$50 = 9B + 3C + D = 12C + 3C + 5 \quad \text{so} \quad C = 3,\ B = 4,\ A = -3.$$

On the other hand, many functions which are not integrable in absolute value on $\mathbb{R}$ are Fourier-transformed to expressions with Dirac δ. For instance,

$$\mathcal{F}(\cos(nt))(\omega) = \sqrt{\frac{\pi}{2}}(\delta(n-\omega) + \delta(n+\omega)),$$

which can easily be seen from the calculation of the Fourier transform of the function $f_\Omega \cos(nx)$ and then be letting Ω approach ∞.

.

We can get the Fourier transform of the sine function in a similar way, we can also take advantage of the fact that the transform of the derivative of this function will differ only by a multiple of the imaginary unit and the variable.

These transforms are a base for Fourier analysis of signals: If a signal is a pure sinusoid of a given frequency, then this is recognized in the Fourier transform as two single-point impulses right at the positive and negative value of the frequency. If the signal is a linear combination of several such pure signals, we obtain the same linear combination of single-point impulses. However, since we always process a signal in a finite time interval only, we actually 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 the information about the original amplitude of the signal.

7.34. Fourier sine a cosine transform. If we apply the Fourier transform to an odd function $f(t)$, i. e., $f(-t) = -f(t)$, the contribution in the integration of the product of $f(t)$ and the function $\cos(\pm\omega t)$ cancels for positive and negative values of t. Thus straight calculation gives

$$\mathcal{F}(f)(\omega) = \frac{-2i}{\sqrt{2\pi}}\int_0^\infty f(t)\sin\omega t\,dt.$$

The resulting function is odd again, hence by the same reason, the inverse transform can determined in a similar way:

$$\tilde{\mathcal{F}}(f)(\omega) = \frac{2i}{\sqrt{2\pi}}\int_0^\infty f(t)\sin\omega t\,dt.$$

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, one can define the *Fourier cosine transform* for even functions:

$$\tilde{f}_c(\omega) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(t)\cos(\omega t)\,dt,$$

$$f(t) = \sqrt{\frac{2}{\pi}}\int_0^\infty \tilde{f}_s(t)\sin\omega t\,dt.$$

7.35. Laplace transforms. The Fourier transform cannot be applied to functions which are not integrable in absolute value over the entire $\mathbb{R}$ (at least, we do not obtain true functions). The so-called *Laplace transform* acts quite similarly as the Fourier one and is flawless in this sense:

$$\mathcal{L}(f)(s) = \bar{f}(s) = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt.$$

Since

$$\frac{s+10}{(s+3)^2}=\frac{s+3+7}{(s+3)^2}=\frac{1}{s+3}+\frac{7}{(s+3)^2},$$

we have

$$\begin{aligned}\mathcal{L}(y)(s) &= \tfrac{-3s+4}{s^2+1}+\tfrac{3}{s+3}+\tfrac{5}{(s+3)^2}+\tfrac{1}{s+3}+\tfrac{7}{(s+3)^2}\\ &= \tfrac{-3s}{s^2+1}+\tfrac{4}{s^2+1}+\tfrac{4}{s+3}+\tfrac{12}{(s+3)^2}.\end{aligned}$$

Hence, using the inverse Laplace transform, we get the solution in the form

$$y(t)=-3\cos t+4\sin t+4e^{-3t}+12te^{-3t}.$$

□

7.48. Find the Laplace transform of Heaviside's function $H(t)$ and shifted Heaviside's function $H_a(t)=H(t-a)$:

$$H(t)=\begin{cases}0 & \text{for } t<0,\\ \frac{1}{2} & \text{for } t=0,\\ 1 & \text{for } t>0.\end{cases}$$

Solution.

$$\begin{aligned}\mathcal{L}(H(t))(s) &= \int_0^\infty H(t)e^{-st}\,\mathrm{d}t=\int_0^\infty e^{-st}\,\mathrm{d}t\\ &= \left[-\frac{e^{-st}}{s}\right]_0^\infty=-\tfrac{1}{s}(0-1)=\tfrac{1}{s},\\ \mathcal{L}(H_a(t))(s) &= \mathcal{L}(H(t-a))(s)=\int_0^\infty H(t-a)e^{-st}\,\mathrm{d}t=\int_a^\infty e^{-st}\,\mathrm{d}t\\ &= \int_0^\infty e^{-s(t+a)}\,\mathrm{d}t=e^{-as}\mathcal{L}(H(t))(s)=\frac{e^{-as}}{s}.\end{aligned}$$

□

7.49. Show that

$$\mathcal{L}(f(t)\cdot H_a(t))(s)=e^{-as}\mathcal{L}(f(t+a))(s) \tag{7.36}$$

Solution.

$$\begin{aligned}\mathcal{L}(f(t)\cdot H_a(t))(s) &= \int_0^\infty f(t)H(t-a)e^{-st}\,\mathrm{d}t=\int_a^\infty f(t)e^{-st}\,\mathrm{d}t\\ &= \int_0^\infty f(t+a)e^{-s(t+a)}\,\mathrm{d}t=e^{-as}\int_0^\infty f(t+a)e^{-st}\,\mathrm{d}t\\ &= e^{-as}\mathcal{L}(f(t+a))(s).\end{aligned}$$

□

7.50. **Solution.** 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 the function $f(t)$ is piecewise continuous:

$$f(t)=\begin{cases}\cos(2t) & \text{for } 0\le t<\pi,\\ 0 & \text{for } t\ge\pi.\end{cases}$$

This problem is a model of undamped oscillation of a spring (excluding friction and other phenomena like non-linearities in the toughness of the spring and so on) which is initiated by an outer force during the initial period only and then ceases.

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.

In an analogous way as in the case of the Fourier transform, we can get the formula for the Laplace transform of a differentiated function for $s>0$ using integration by parts:

$$\begin{aligned}\mathcal{L}(f'(t))(s) &= \int_0^\infty f'(t)\,\mathrm{e}^{-st}\,dt\\ &= [f(t)\,\mathrm{e}^{-st}]_0^\infty+s\int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt\\ &= -f(0)+s\mathcal{L}(f)(s).\end{aligned}$$

The properties of the Laplace transform and many more transforms used especially in technical practice can be found in specialized literature.

4. Discrete transforms

The Fourier analysis of signals mentioned in the previous paragraph used to be realized by special analog circuits in radio technology, for instance. Nowadays, we work only with discrete data when processing signals by computer circuits. Let us assume that there is a fixed (tiny) *sample interval* τ given in a (discrete) time variable and that our signal repeats with period $N\tau$ (for a very large natural number N), which is the maximal period which can be represented in our discrete model.

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, making use of the previous examples 7 and 8, the right-hand sides to the calculation of the Laplace transform

$$\begin{aligned} s^2\mathcal{L}(y) + 1 + 4\mathcal{L}(y) &= \mathcal{L}(\cos(2t)(u(t) - u_\pi(t))) \\ &= \mathcal{L}(\cos(2t)\cdot u(t)) - \mathcal{L}(\cos(2t)\cdot u_\pi(t)) \\ &= \mathcal{L}(\cos(2t)) - e^{-\pi s}\mathcal{L}(\cos(2(t+\pi)) \\ &= (1 - e^{-\pi s})\frac{s}{s^2+4}. \end{aligned}$$

Hence,

$$\mathcal{L}(y) = -\frac{1}{s^2+4} + (1 - e^{-\pi s})\frac{s}{(s^2+4)^2}.$$

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 (‖7.36‖),

$$\begin{aligned} \mathcal{L}^{-1}\left(e^{-\pi s}\frac{s}{(s^2+4)^2}\right) &= \tfrac{1}{4}\mathcal{L}^{-1}(e^{-\pi s}\mathcal{L}(t\sin(2t))) \\ &= (t-\pi)\sin(2(t-\pi))\cdot H_\pi(t). \end{aligned}$$

Since Heaviside's function is zero for $t < \pi$ and equals 1 for $t > \pi$, we get the solution in the form

$$y(t) = \begin{cases} -\frac{1}{2}\sin(2t) + \frac{1}{4}t\sin(2t) & \text{for } 0 \le t < \pi \\ \frac{\pi-2}{4}\sin(2t) & \text{for } t \ge \pi \end{cases}$$

□

7.51. Find a function $y(t)$ satisfying the differential equation

$$y''(t) = \cos(\pi t) - y(t), \qquad t \in (0, +\infty)$$

and the initial conditions $y(0) = c_1$, $y'(0) = c_2$.

Solution. First, let us emphasize that from the theory of ordinary differential equations, it follows that this problem has a unique solution. Further, let us remind that

$$\mathcal{L}\left(f''\right)(s) = s^2\mathcal{L}(f)(s) - s\lim_{t\to 0+} f(t) - \lim_{t\to 0+} f'(t)$$

and

$$\mathcal{L}(\cos(bt))(s) = \tfrac{s}{s^2+b^2}, \qquad b \in \mathbb{R}.$$

Applying the Laplace transform to the given differential equation thus gives

$$s^2\mathcal{L}(y)(s) - sc_1 - c_2 = \tfrac{s}{s^2+\pi^2} - \mathcal{L}(y)(s),$$

i. e.,

$$\mathcal{L}(y)(s) = \frac{s}{\left(s^2+1\right)\left(s^2+\pi^2\right)} + \frac{c_1 s}{s^2+1} + \frac{c_2}{s^2+1}. \tag{7.37}$$

It suffices to find a function y satisfying (‖8.12‖). Partial fraction decomposition gives

$$\tfrac{s}{(s^2+1)(s^2+\pi^2)} = \tfrac{1}{\pi^2-1}\left(\tfrac{s}{s^2+1} - \tfrac{s}{s^2+\pi^2}\right).$$

Therefore, from the expression of $\mathcal{L}(\cos(bt))(s)$ mentioned above and the proved equality

$$\mathcal{L}(\sin t)(s) = \tfrac{1}{s^2+1},$$

we already obtain the wanted solution

$$y(t) = \tfrac{1}{\pi^2-1}(\cos t - \cos(\pi t)) + c_1 \cos t + c_2 \sin t.$$

□

7.52. 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 initial conditions $x(0) = 0$, $y(0) = 0$, $x'(0) = 1$, $y'(0) = 0$.

Solution. Once again, we apply the Laplace transform. Together with

$$\mathcal{L}\left(\mathrm{e}^{\pm t}\right)(s) = \tfrac{1}{s\mp 1},$$

this transforms the first equation to

$$s^2\mathcal{L}(x)(s) - s\lim_{t\to 0+} x(t) - \lim_{t\to 0+} x'(t) + s\mathcal{L}(x)(s) - \lim_{t\to 0+} x(t) =$$
$$\mathcal{L}(y)(s) - \left(s^2\mathcal{L}(y)(s) - s\lim_{t\to 0+} y(t) - \lim_{t\to 0+} y'(t)\right) + \tfrac{1}{s-1}$$

and the second one to

$$s\mathcal{L}(x)(s) - \lim_{t\to 0+} x(t) + 2\mathcal{L}(x)(s) =$$
$$-\mathcal{L}(y)(s) + s\mathcal{L}(y)(s) - \lim_{t\to 0+} y(t) + \tfrac{1}{s+1}.$$

If we enumerate the limits (according to the initial conditions), we get the linear equations

$$s^2\mathcal{L}(x)(s) - 1 + s\mathcal{L}(x)(s) = \mathcal{L}(y)(s) - s^2\mathcal{L}(y)(s) + \tfrac{1}{s-1}$$

and

$$s\mathcal{L}(x)(s) + 2\mathcal{L}(x)(s) = -\mathcal{L}(y)(s) + s\mathcal{L}(y)(s) + \tfrac{1}{s+1}$$

with a unique solution

$$\mathcal{L}(x)(s) = \tfrac{2s-1}{2(s-1)(s+1)^2}, \quad \mathcal{L}(y)(s) = \tfrac{3s}{2\left(s^2-1\right)^2}.$$

Once again, we use partial fraction decomposition, obtaining

$$\mathcal{L}(x)(s) = \tfrac{1}{8}\tfrac{1}{s-1} + \tfrac{3}{4}\tfrac{1}{(s+1)^2} - \tfrac{1}{8}\tfrac{1}{s+1} = \tfrac{3}{4}\tfrac{1}{(s+1)^2} + \tfrac{1}{4}\tfrac{1}{s^2-1}.$$

Since we have already calculated that

$$\mathcal{L}\left(t\,\mathrm{e}^{-t}\right)(s) = \tfrac{1}{(s+1)^2}, \quad \mathcal{L}(\sinh t)(s) = \tfrac{1}{s^2-1},$$
$$\mathcal{L}(t\sinh t)(s) = \tfrac{2s}{\left(s^2-1\right)^2},$$

we get

$$x(t) = \tfrac{3}{4}t\,\mathrm{e}^{-t} + \tfrac{1}{4}\sinh t, \quad y(t) = \tfrac{3}{4}t\sinh t.$$

The reader can verify that these functions x and y are really the wanted solution. We strongly recommend to perform the verification (for instance for the reason that the Laplace transforms of the functions $y = \mathrm{e}^t$, $y = \sinh t$ and $y = t\sinh t$ were obtained only for $s > 1$). □

7.53. Find the solution to the following system of differential equations:

$$\begin{aligned} x'(t) &= -2x(t) + 3y(t) + 3t^2, \\ y'(t) &= -4x(t) + 5y(t) + e^t, \quad x(0) = 1,\ y(0) = -1 \end{aligned}$$

Solution.

$$\begin{aligned} \mathcal{L}(x')(s) &= \mathcal{L}(-2x + 3y + 3t^2)(s), \\ \mathcal{L}(y')(s) &= \mathcal{L}(-4x + 5y + e^t)(s). \end{aligned}$$

The left-hand sides can be written using (‖8.11‖), and the right-hand ones can be rewritten thanks to linearity of the operator $\mathcal{L}$. Since $\mathcal{L}(3t^2)(s) = \frac{6}{s^3}$ and $\mathcal{L}(e^t)(s) = \frac{1}{s-1}$, we get the system of linear equations

$$\begin{aligned} s\mathcal{L}(x)(s) - 1 &= -2\mathcal{L}(x)(s) + 3\mathcal{L}(y)(s) + \tfrac{6}{s^3}, \\ s\mathcal{L}(y)(s) + 1 &= -4\mathcal{L}(x)(s) + 5\mathcal{L}(y)(s) + \tfrac{1}{s-1}. \end{aligned}$$

After rearrangements, we get $\mathbf{A}(s)\hat{\mathbf{x}}(s) = \mathbf{b}(s)$ in matrices, where we denoted

$\mathbf{A}(s) = \begin{pmatrix} s+2 & -3 \\ 4 & s-5 \end{pmatrix}$, $\hat{\mathbf{x}}(s) = \begin{pmatrix} \mathcal{L}(x)(s) \\ \mathcal{L}(y)(s) \end{pmatrix}$ and $\mathbf{b}(s) = \begin{pmatrix} 1 + \frac{6}{s^3} \\ -1 + \frac{1}{s-1} \end{pmatrix}$.

Cramer's rule says that

$$\mathcal{L}(x)(s) = \tfrac{|\mathbf{A}_1|}{|\mathbf{A}|}, \quad \mathcal{L}(y)(s) = \tfrac{|\mathbf{A}_2|}{|\mathbf{A}|}, \text{ where}$$

$$\begin{aligned} |\mathbf{A}| &= \begin{vmatrix} s+2 & -3 \\ 4 & s-5 \end{vmatrix} = s^2 - 3s + 2, \\ |\mathbf{A}_1| &= \begin{vmatrix} 1+\frac{6}{s^3} & -3 \\ -1+\frac{1}{s-1} & s-5 \end{vmatrix} = (s-5)(1+\tfrac{6}{s^3}) + 3(-1+\tfrac{1}{s-1}) \\ |\mathbf{A}_2| &= \begin{vmatrix} s+2 & 1+\frac{6}{s^3} \\ 4 & -1+\frac{1}{s-1} \end{vmatrix} = (s+2)(-1+\tfrac{1}{s-1}) - 4 - \tfrac{24}{s^3}. \end{aligned}$$

Hence

$$\begin{aligned} \mathcal{L}(x)(s) &= \frac{1}{(s-1)(s-2)}\left(\frac{(s-5)(s^3+6)}{s^3} - 3\frac{s-2}{s-1}\right), \\ \mathcal{L}(y)(s) &= \frac{1}{(s-1)(s-2)}\left(\frac{(s+2)(2-s)}{s-1} - \frac{4s^3+24}{s^3}\right). \end{aligned}$$

Using partial fraction decomposition, we can express the Laplace images of the solution

$$\begin{aligned} \mathcal{L}(x)(s) &= -\tfrac{39}{2s^2} - \tfrac{3}{(s-1)^2} + \tfrac{28}{s-1} - \tfrac{21}{4(s-2)} - \tfrac{15}{s^3} - \tfrac{87}{4s}, \\ \mathcal{L}(x)(s) &= -\tfrac{18}{s^2} - \tfrac{3}{(s-1)^2} + \tfrac{27}{s-1} - \tfrac{7}{s-2} - \tfrac{12}{s^3} - \tfrac{21}{s} \end{aligned}$$

and then we arrive at the solutions of Cauchy's problem with the inverse transformation:

$$\begin{aligned} x(t) &= -\tfrac{39}{2}t - 3te^t + 28e^t - \tfrac{21}{4}e^{2t} - \tfrac{15}{2}t^2 - \tfrac{87}{4}, \\ y(t) &= -18t - 3te^t + 27e^t - 7e^{2t} - 6t^2 - 21. \end{aligned}$$

□

7.54. Discrete cosine transform. The fundamental feature of JPEG compression of data is the so-called discrete cosine transform. That is given by an orthogonal matrix $C = (c_{kl})_{k,l=1}^n$ defined as follows:

$$c_{kl} = \alpha_{kl}\cos\left(\frac{(2k-1)(l-1)\pi}{2n}\right)$$

where $\alpha_{k1} = \frac{1}{\sqrt{n}}$ $\alpha_{kl} = \sqrt{\frac{2}{n}}$ for $l > 1$. The vector representing the data is then decomposed orthogonally and some basis vectors (columns of the matrix C) are dropped. This produces a reduction of data reasonably approximating the original set. The inverse transform is easy. Since C is orthogonal, it is given by multiplication by the transposed matrix.

Show that for $n = 2$, the matrix C equals $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and that it is orthogonal. Calculate the orthogonal decomposition of vector (3, 4) with respect to the basis formed by the columns of the matrix and determine the eigenvalues and eigenvectors.

Solution. Let us calculate

$$CC^T = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\cdot\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix} = 1.$$

Thus, the matrix C is really orthogonal and its columns create an orthonormal basis $e_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, $e_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. The coefficients of the orthogonal decomposition of the vector $u = (3, 4)$ can be obtained easily by applying the transposed matrix

$$C^T u = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}3\\ 4\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}7\\ -1\end{pmatrix}$$

Therefore, the orthogonal decomposition has the following form:

$$\begin{pmatrix}3\\ 4\end{pmatrix} = \frac{7}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix} - \frac{1}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}}\end{pmatrix}$$

The characteristic polynomial of the matrix C is $(\lambda+\frac{1}{\sqrt{2}})(\lambda-\frac{1}{\sqrt{2}})-\frac{1}{2} = 0$, so the eigenvalues are $\lambda_{1,2} = \pm 1$ (an orthogonal matrix cannot have any others). The corresponding eigenvectors are determined by the respective equations

$$(\tfrac{1}{\sqrt{2}} - 1)x + \tfrac{1}{\sqrt{2}}y = 0, \quad (\tfrac{1}{\sqrt{2}} + 1)x + \tfrac{1}{\sqrt{2}}y = 0.$$

So these are, for instance, the vectors $(\frac{1}{\sqrt{2}}, 1 - \frac{1}{\sqrt{2}})$, $(\frac{1}{\sqrt{2}}, -1 - \frac{1}{\sqrt{2}})$ (which are orthogonal automatically). □

Remark. Try to draw a picture of the action of the mapping determined by the matrix A on some vector in the plane.

7.55. Discrete cosine transform 2. Show that the symmetric matrix

$$\frac{1}{2}\begin{pmatrix}0 & 1 & \dots & 0 & 0\\ 1 & 0 & \dots & 0 & 0\\ \vdots & & & & \vdots\\ 0 & 0 & \dots & 0 & 1\\ 0 & 0 & \dots & 1 & 0\end{pmatrix}$$

has eigenvalues $\lambda_l = \cos\varphi_l$, where $\varphi_l = \frac{l\pi}{n+1}$ with $1 \le l \le n$, and that the corresponding eigenvectors $\sqrt{\frac{2}{n+1}}(\sin\varphi_l, \sin 2\varphi_l, \vdots \sin n\varphi_l)$ form an orthonormal basis.

Solution. First, we calculate the k-th coordinate of the vector

$$\frac{1}{2}\sqrt{\frac{2}{n+1}}\begin{pmatrix}0 & 1 & \dots & 0 & 0\\ 1 & 0 & \dots & 0 & 0\\ \vdots & & & & \vdots\\ 0 & 0 & \dots & 0 & 1\\ 0 & 0 & \dots & 1 & 0\end{pmatrix}\begin{pmatrix}\sin\varphi_l\\ \sin 2\varphi_l\\ \vdots\\ \sin n\varphi_l\end{pmatrix}$$

Using the sine sum formula, we obtain

$$\frac{1}{2}\sqrt{\frac{2}{n+1}}(\sin(k-1)\varphi_l + \sin(k+1)\varphi_l) = \sqrt{\frac{2}{n+1}}\sin k\varphi_l\cos\varphi_l,$$

so the given vector is really an eigenvector corresponding to eigenvalue $\cos\varphi_l$. Since there are n distinct eigenvalues (which is the dimension), these eigenvectors form a basis. It only remains to verify that the eigenvectors are orthogonal and normed. □

E. Additional exercises to the whole chapter

7.56. Expand the function $\sin^2(x)$ on the interval $[-\pi, \pi]$ into the Fourier series. ○

7.57. Expand the function $\cos^2(x)$ on the interval $[-\pi, \pi]$ into the Fourier series. ○

7.58. Determine the convolution of the functions f_1 and f_2, where

$$\begin{aligned} f_1 &= \begin{cases} 1 & \text{for } x \in [-1, 0] \\ 0 & \text{otherwise} \end{cases} \\ f_2 &= \begin{cases} x & \text{for } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

○

Key to the exercises

7.4. 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.

7.5. $\cos(x)$, $\frac{4}{\pi}\cos(x) + x$. The projection does not change the function $\frac{1}{3}\cos(x)$ since it already lies in the space.

7.34.

$$f_1 * f_2(t) = \begin{cases} t - \frac{t^2}{2} + 4 & \text{for } t \in [-2, -1] \\ 1 - t + \frac{1}{2} & \text{for } t \in [-1, 1] \\ \frac{t^2}{2} - 2t + 2 & \text{for } t \in [1, 2] \\ 0 & \text{otherwise} \end{cases}$$

7.56. $\frac{1}{2} - \frac{1}{2}\cos(2x)$.

7.57. $\frac{1}{2} + \frac{1}{2}\cos(2x)$.

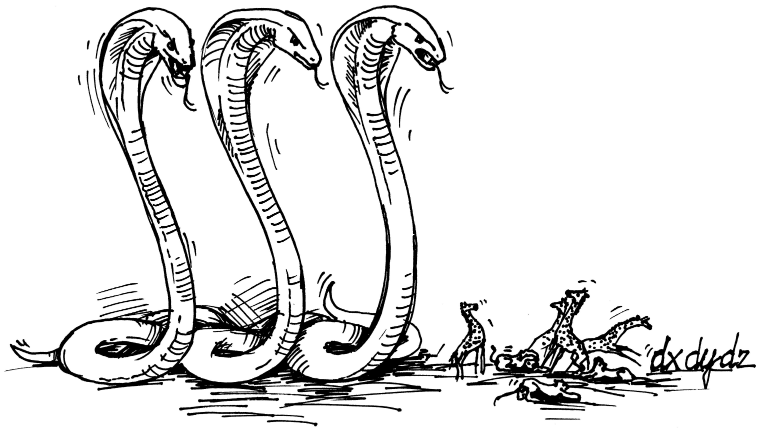
7.58. Determine the convolution of the functions f_1 and f_2, where

$$f_1 * f_2(t) \quad = \quad \begin{cases} \frac{(t+1)^2}{2} & \text{for } t \in [-1, 0] \\ \frac{1-t^2}{2} & \text{for } t \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

CHAPTER 8

Continuous models with more variables

one variable is not enough?
– never mind, just recall vectors!



A. Multivariate functions

8.1. Determine the domain of the function $\mathbb{R}^2 \to \mathbb{R}$ which is given by the following formula:

a)
$$\frac{xy}{y(x^3+x^2+x+1)},$$

b)
$$\ln(x^2-y^2),$$

c)
$$\ln(-x^2-y^2),$$

d)
$$\arcsin(2\,\mathrm{sgn}(\chi_{\mathbb{Q}}(x)),$$

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 if and only if 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.

At the very beginning of our journey through the mathematical countryside, we have seen that it is not difficult to work with more parameters simultaneously since vectors can be manipulated as easily as scalars. We only have to think things out well. Now, once again, we will deal with situations when the mathematically expressed relations depend on more (yet still finitely many) parameters. We will see that there is no need of brand new ideas; it will often do to reduce the problems we encounter to the ones we are able to solve.

At the same time, we can finally return to the discussion of situations when the function values are described in terms of instantaneous changes – i. e., we will stop for a while to look at ordinary and partial differential equations. At the very end, we will introduce the so-called variation problems.

As usual, we will try to comment on the discrete variants of our approaches or problems on the fly.

1. Functions and mappings on $\mathbb{R}^n$

8.1. Multivariate functions. For the modeling of processes (or objects in graphics) in practice, we can seldom do with functions $\mathbb{R} \to \mathbb{R}$ of one variable. At least, functions dependent on parameters are necessary, and the dependence of the change of the results on the parameters is often more important than the result itself. Therefore, we will consider the functions

$$f(x_1, x_2, \dots, x_n) : \mathbb{R}^n \to \mathbb{R},$$

and we will try to extend our methods for monitoring the values and their changes for this situation. We call them *functions of more variables* or, more compactly, *multivariate functions*.

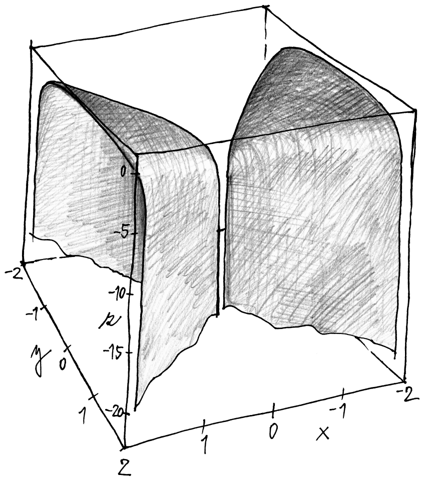
We will often work with the cases $n = 2$ or $n = 3$ so that the concepts being introduced would be easier to understand, and we will, in these cases, use the letters x, y, z instead of numbered variables. This means that a function f defined in the "plane" $\mathbb{R}^2$ will be denoted

$$f : \mathbb{R}^2 \ni (x,y) \mapsto f(x,y) \in \mathbb{R},$$

and, similarly, in the "space" $\mathbb{R}^3$

$$f : \mathbb{R}^3 \ni (x,y,z) \mapsto f(x,y,z) \in \mathbb{R}.$$

Just like in the case of univariate functions, we talk about the domain $A \subset \mathbb{R}^n$ on which the function in question is defined. 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.



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 $(x, y) \in R^2$ 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, the argument of the natural logarithm must be positive, and the argument of the arc sine must be from $[-1, 1]$. □

B. The topology of E_n

8.2. A known fact about the space E_n is that the shortest path between a pair of points is a line segment. However, many more 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.

8.3. Show that every non-empty proper subset of E_n has a boundary point (which need not lie in it).

Solution. Let $U \subset E_n$ be a non-empty subset with no boundary point. Consider a point $X \in U$, a point $Y \in U' := E_n \setminus U$, and the line segment $XY \subset E_n$. Intuitively, going from X to Y along this segment, we must once get from U to U', and this can happen only at a boundary point (everyone who has ever been to a foreign country is surely well acquainted with this fact). Formally, let A be the point of XY for which $|XA| = \sup\{|XZ|, XZ \in U\}$ (apparently, there is exactly one such point on the segment XY). This point is a boundary point of U: it follows from the definition of A that any line segment XB (with $B \in XA$) is contained in U; in particular, $B \in U$. However, if there were a neighborhood of A contained in U, then there would exist a part of the line segment XY longer than XA which would be contained in U, which contradicts the definition of the point A. Therefore, any neighborhood of the point A contains a point from U as well as a point from $E_n \setminus U$. □

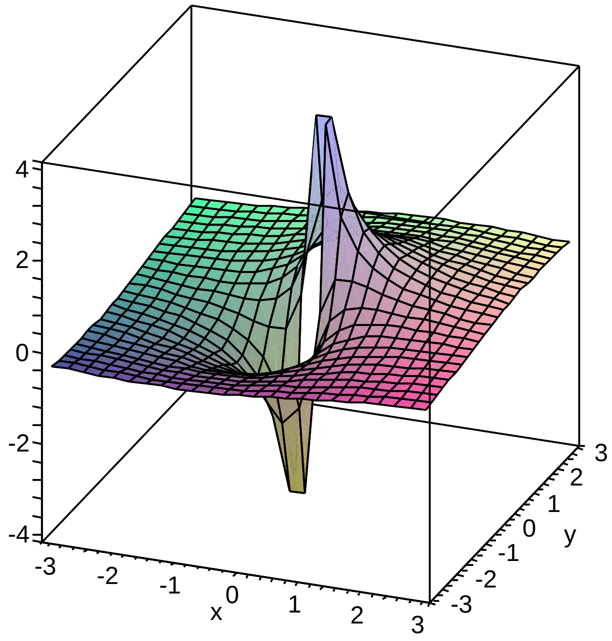
It is also useful to consider the *graph* of a multivariate function, i. e., a subset $G_f \subset \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$, defined by

$$G_f = \{(x_1, \dots, x_n, f(x_1, \dots, x_n));\ (x_1, \dots, x_n) \in A\},$$

where A is the domain of f. For instance, the graph of the function defined in the plane by the formula

$$f(x, y) = \frac{x + y}{x^2 + y^2}$$

is quite a nice surface, caught in the picture. 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, we used fixed *coordinates* in the plane. If we fix the value of either of the coordinates, only one variable remains. Fixing the value of x, for example, we get 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, with which we have already worked, namely in chapter six (see 6.14). The images of the curves for some fixed values of the coordinates x and y are depicted by lines in the picture.

The curves $c : \mathbb{R} \to \mathbb{R}^n$ are, besides multivariate functions, the easiest examples of *mappings* $F : \mathbb{R}^m \to \mathbb{R}^n$, which we will get to shortly.

In the case of functions of one variable, we built the entire differential and integral calculi upon the concepts of convergence, open neighborhoods, continuity, and so on. In the second part of chapter seven, these concepts were generalized for the so-called metric spaces, rather than only for the Euclidean spaces $\mathbb{R}^n$.

Before going on with the following paragraphs, it is appropriate to recall these parts, or to look for the concepts and results there when necessary. We present a bit of a summary here.

8.4. 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 ‖8.3‖ 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. □

8.5. 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., $E_n = \underset{i\in I}{\cup} U_i$, 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 ‖8.4‖. □

8.6. 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 in general. 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 $(0, 1)$. □

8.7. 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 $\{1\}$. □

8.8. 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.

8.2. Euclidean spaces. A Euclidean space E_n is perceived as a set of points in $\mathbb{R}^n$ without any choice of coordinates, and its direction space $\mathbb{R}^n$ is considered to be the vector space of all increases that can be added to the points of the space E_n.

Moreover, a standard scalar product

$$u \cdot v = \sum_{i=1}^{n} x_i y_i,$$

is selected on $\mathbb{R}^n$, where $u = (x_1, \dots, x_n)$ and $v = (y_1, \dots, y_n)$ are arbitrary vectors. This gives a *metric* on E_n, i. e., a function describing the distance $\|P - Q\|$ of 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 of the points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ is given by

$$\|P_1 - P_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2.$$

Metrics defined in this manner satisfy the triangle inequality for every triple of points P, Q, R:

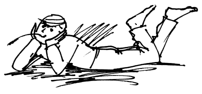
$$\|P - R\| = \|(P - Q) + (Q - R)\| \le \|(P - Q)\| + \|(Q - R)\|,$$

see 3.25(1) in geometry, or the axioms of a metric in 7.12, or the same inequality (5.4) for scalars. The concepts defined for real and complex scalars and discussed for metric spaces in detail thus can be carried over (extended) with no problem for the points P_i of any Euclidean space:

The topology of Euclidean spaces

- *a Cauchy sequence*: a sequence of points P_i such that for every fixed $\varepsilon > 0$, $\|P_i - P_j\| < \varepsilon$ holds for all indeces but for finitely many exceptional values i, j;
- *a convergent sequence*: a sequence of points P_i converges to a point P iff for every fixed $\varepsilon > 0$, $\|P_i - P\| < \varepsilon$ holds for all but finitely many indeces i; the point P is then called the *limit* of the sequence P_i;
- *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;
- *a closed set*: contains all of its limit points;
- *an open set*: its complement is closed;
- *an open δ–neighborhood of a point P*: the set $\mathcal{O}_\delta(P) = \{Q \in E_n;\ \|P - Q\| < \delta\}$, $\delta \in \mathbb{R}$, $\delta > 0$;
- *a boundary point P of a set A*: every δ–neighborhood of P has non-empty intersection with both A and the complement $E_n \setminus A$;
- *an interior point P of a set A*: there exists a δ–neighborhood of P which lies inside A;
- *a bounded set*: lies inside some δ–neighborhood of one of its points (for a sufficiently large δ);
- *a compact set*: both closed and bounded.

The reader should make an appropriate effort to read the paragraphs 3.25, 5.14–5.17, 7.14–7.16, and 7.22 as well as try to think out/recall the definitions and connections of all these concepts.



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.) □

C. Tangent lines, tangent planes, graphs of multivariate functions

8.9. 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 8.4, 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 8.4 starts at the point $(0, 0, 0)$ (in other words, $C(0) = (0, 0, 0)$). In the affine space $\mathbb{R}^n$, we can move it so that it starts at an arbitrary point, and this does not change its derivative (this is performed by adding a constant to every component in the parametric equation of the curve). Therefore, up to the movement, this integral curve is determined uniquely (nothing else than constants can be added to the components without changing the derivative). When we integrate the curve of acceleration, we get the curve of velocity $(-\sin t, \cos t - 1, 0)$. Considering the initial velocity as well, we obtain the velocity curve of the car: $(-\sin t, \cos t, 1)$ (we shifted the curve of the vector $(0, 1, 1)$, i. e., so that now the velocity curve at time $t = 0$ agrees with the given initial velocity). Further integration leads to the curve $(\cos t - 1, \sin t, t)$. Shifting this of the vector $(1, 0, 0)$ then fits with the initial position of the car. Therefore, the car moves along the curve $[\cos t, \sin t, t]$ (this curve is called a helix). □

8.10. 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:

$$\begin{aligned} x &= c_1'(1)s + c_1(1) = t + 1, \\ y &= c_2'(1)s + c_2(1) = 2t + 1, \\ z &= c_3'(1)s + c_3(1) = 3t + 1. \end{aligned}$$

In order to get the implicit equations (which are not given canonically), we eliminate the parameter t, thereby obtaining:

$$\begin{aligned} 2x - y = 1, \\ 3x - z = 2. \end{aligned}$$

□

8.11. 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 $\mathbb{R}^2$; $\text{In}(x + y)$ is a differentiable function on the set of points with $x > y$ (an open half-plane, i. e., without the

It should be clear straight from the definitions that sequences of points P_i have the properties mentioned in the first and second items 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 5.12 that every Cauchy sequence of points in E_n is convergent. Especially, E_n is always a complete metric space.

8.3. Compact sets. Our games with open, closed, or compact sets could seem useless in the case of the real line E_1 since in the end, we almost always talked about intervals.

In the case of metric spaces in the second part of chapter seven, it probably was, on the other hand, too complicated. However, the same approach is quite 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 like in the case of E_1, we define the open cover of a set (i. e., a system of open sets containing the given set), and the Theorem 5.17 holds as well (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–neighborhoods,
(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 5.17 can be reused without changes in the case of propositions (1)–(3), yet now the concepts have to be perceived in a different way, and the "open intervals" are substituted with multidimensional δ–neighborhoods of appropriate points.

However, the proof of the fourth and fifth propositions has to be adjusted properly. Therefore, it is a good idea to go through the proof of the corresponding propositions for general metric spaces in 7.22 while noticing the parts which can be simplified for Euclidean spaces. □

8.4. Curves in E_n. Almost all of our 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 we used only the triangle inequality valid for the magnitudes of the real and complex numbers. We already noticed back then that this argument can be carried over to any functions of a real variable with values in a Euclidean space $\mathbb{R}^n$, and we introduced several tools for the work with curves in paragraphs 6.14–6.17.

Therefore, let us remind that for every (parametrized) curve[1], i. e., a mapping $c : \mathbb{R} \to \mathbb{R}^n$ in an n–dimensional space, we can work with the concepts which simply extend our reasonings from the univariate functions:

- *a limit:* $\lim_{t \to t_0} c(t) \in \mathbb{R}^n$

[1] In geometry, one often makes a distinction between a curve as a subset of E_n and its parametrization $\mathbb{R} \to \mathbb{R}^n$. When we say the word "curve", we will exclusively mean the parametrized curve.

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 $x_1, \dots, x_n$ with respect to the variable x_1 will be denoted by both $\frac{\partial f}{\partial x_1}$ and the shorter expression f_{x_1}. In the exercise part of the book, we will rather keep to the latter notation. On the other hand, the notation $\frac{\partial f}{\partial x_1}$ better catches the fact that this is a derivative of f in the direction of the vector field $\frac{\partial}{\partial x_1}$ (you will learn what a vector field is in paragraph 8.34).

8.12. 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 $y = 0$. □

8.13. 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 8.5). The second one is to use the differential of the function; see 8.6 and theorem 8.7. Since the given function is a polynomial, it is differentiable on the whole $\mathbb{R}^3$.

Let us go from the definition:

$$\begin{aligned} f_v(x, y, z) &= \lim_{t\to 0}\frac{1}{t}[f(x + 3t, y + 2t, z - t) - f(x, y, z)] = \\ &= \lim_{t\to 0}\frac{1}{t}[(x + 3t)^2(y + 2t)(z - t) - x^2yz] = \\ &= \lim_{t\to 0}\frac{1}{t}[t(6xyz + 2x^2z - x^2y) + t^2(\dots)] = \\ &= 6xyz + 2x^2z - x^2y. \end{aligned}$$

We have thus derived the derivative in the direction of the vector $(3, 2, -1)$ as a function of three real variables which determine the point at which we are interested in the value of the derivative. Evaluating this for the desired point thus leads to $f_v(1, -1, 2) = -7$.

In order to compute the directional derivative from the differential of the function, we first have to determine the partial derivatives of the function:

$$f_x = 2xyz,\ f_y = x^2z,\ f_z = x^2y.$$

It follows from the note beyond theorem 8.7 that we can express $f_v(1, -1, 2) = 3f_x(1, -1, 2) + 2f_y(1, -1, 2)+$

- *a derivative:* $c'(t_0) = \lim_{t\to t_0}\frac{1}{|t-t_0|}\cdot(c(t) - c(t_0)) \in \mathbb{R}^n$
- *an integral:* $\int_a^b c(t)dt \in \mathbb{R}^n$.

We can also notice 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 direction space!). In the case of an integral, we will have to consider curves in the vector space $\mathbb{R}^n$. The reason for this can be seen even in the case of one dimension, where we need to know the origin to be able to see the "area under the graph of a function".

Once again, it is apparent straight from the definition that limits, derivatives, and integrals can be calculated by particular n-coordinate components in $\mathbb{R}^n$, and their existence can be determined in the same way.

We can also directly formulate the analogy of the connection between the Riemann integral and the antiderivative for curves (see 6.25):

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 it holds that $C'(t) = c(t)$ for all values $t \in [a, b]$.

It is worse with the mean value theorem and, in general, with Taylor's theorem, see 5.38 and 6.4. We can apply them 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, we get the existence of numbers t_i such that

$$c_i(b) - c_i(a) = (b - a)\cdot c_i'(t_i).$$

However, these numbers t_i will be distinct in general, so we cannot express the difference vector of the marginal points $c(b) - c(a)$ as a multiple of the derivative of the curve at a single point. For example, in the plane E_2, we thus get for the differentiable curve $c(t) = (x(t), y(t))$ that

$$\begin{aligned} c(b) - c(a) &= (x'(\xi)(b - a), y'(\eta)(b - a)) \\ &= (b - a)\cdot(x'(\xi), y'(\eta)) \end{aligned}$$

for two (different, in general) values $\xi, \eta \in [a, b]$. However, this reasoning is still sufficient for the following bound:

Lemma. *If c is a curve in E_n with continuous derivative on a compact interval $[a, b]$, then we have for all $a \leq s \leq t \leq b$ that*

$$\|c(t) - c(s)\| \leq \sqrt{n}(\max_{r\in[a,b]}\|c'(r)\|)\cdot|t - s|.$$

Proof. Direct application of the mean value theorem gives for appropriate points r_i inside the interval $[s, t]$ the following:

$$\begin{aligned} \|c(t) - c(s)\|^2 &= \sum_{i=1}^n(c_i(t) - c_i(s))^2 \leq \sum_{i=1}^n(c_i'(r_i)(t - s))^2 \\ &\leq (t - s)^2\sum_{i=1}^n \max_{r\in[s,t]} c_i'(r)^2 \\ &\leq n(\max_{r\in[s,t],\ i=1,\dots,n}|c_i'(r)|)^2(t - s)^2 \\ &\leq n\max_{r\in[s,t]}\|c'(r)\|^2(t - s)^2. \end{aligned}$$

$$+ (-1) f_z(1,-1,2) =$$
$$= 3\cdot(-4) + 2\cdot 2 + (-1)\cdot(-1) = -7. \qquad \square$$

8.14. Determine the derivative of the function $f : \mathbb{R}^3 \to R$, $f(x,y,z) = \frac{\cos(x^2 y)}{z}$ 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 ||8.11||). We can determine the value of the examined directional derivative by 8.6, using partial derivatives.

First, we determine the partial derivatives of the given function (as we have already mentioned in exercise ||8.12||, 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):

$$f_x = -\frac{2xy\sin(x^2 y)}{z},\ f_y = -\frac{x^2 \sin(x^2 y)}{z},\ f_z = -\frac{\cos(x^2 y)}{z^2}.$$

Evaluating this expression at the particular values, we obtain

$$f_x(0,0,2) + 2\cdot f_y(0,0,2) + 3\cdot f_z(0,0,2) =$$
$$1\cdot 0 + 2\cdot 0 + 3\cdot\left(-\frac{1}{4}\right) = -\frac{3}{4}.$$

□

8.15. 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 8.4, 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 $v_1, \dots, v_n$ 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 ||**??**||). The value $f_v(x)$ can be interpreted as the scalar product of the vectors $(f_{x_1}, \dots, f_{x_n})$ and $(v_1, \dots, v_n)$. And this product is maximal if the vectors are of the same direction. The vector v can thus be obtained by normalizing the vector $(f_{x_1}, \dots, f_{x_n})$. In general, we say the the function grows maximally in the direction $(f_{x_1}, \dots, f_{x_n})$. Then, this vector is called the gradient of the function f. In paragraph 8.19, we will recall this idea and go into further details. □

8.16. 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

$$\begin{aligned} z &= f\left(1, \frac{1}{e}\right) + f_x\left(1, \frac{1}{e}\right)(x-1) + f_y\left(1, \frac{1}{e}\right)\left(y - \frac{1}{e}\right) \\ &= -1 - x + ey. \end{aligned}$$

The given point does not satisfy this equation, so it does not lie in the tangent plane. □

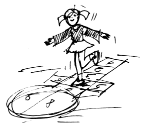
□

An important concept is the one of a *tangent vector* to a curve $c : \mathbb{R} \to E_n$ at a point $c(t_0) \in E_n$, which is defined as the vector in the direction $\mathbb{R}^n$ given by the derivative $c'(t_0) \in \mathbb{R}^n$. The straight line T given parametrically by

$$T : \quad c(t_0) + t\cdot c'(t_0)$$

is called the *tangent line to the curve c* at the point t_0. Unlike the tangent vector, the tangent line T, being a non-parametrized line, is independent of the parametrization of the curve c since thanks to the chain rule, changing the parametrization leads to the same tangent vector, up to multiple.

8.5. Partial derivatives. For every function $f : \mathbb{R}^n \to \mathbb{R}$ and an arbitrary curve $c : \mathbb{R} \to \mathbb{R}^n$, we can consider their composition $(f \circ c)(t) : \mathbb{R} \to \mathbb{R}$. This composite function $F \circ c$ expresses the behavior of the function f along the curve c. The simplest case is when we use straight lines.



DIRECTIONAL AND PARTIAL DERIVATIVES

Definition. We say that $f : \mathbb{R}^n \to \mathbb{R}$ has *derivative in the direction of a vector* $v \in \mathbb{R}^n$ at a point $x \in E_n$ iff the derivative $d_v f(x)$ of the composite mapping $t \mapsto f(x+tv)$ at the point $t = 0$, i. e.,

$$d_v f(x) = \lim_{t\to 0} \frac{1}{t}(f(x+tv) - f(x)).$$

The value $d_v f$ is also called a *directional derivative*.

The special choice of the lines in the direction of the axes of the coordinate system yields the so-called *partial derivatives of the function f*, which are denoted by $\frac{\partial f}{\partial x_i}$, $i = 1, \dots, n$, or (without referring to the function) as operations $\frac{\partial}{\partial x_i}$.

For functions in the plane, we thus get

$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= \lim_{t\to 0} \frac{1}{t}(f(x+t,y) - f(x,y)) \\ \frac{\partial}{\partial y} f(x,y) &= \lim_{t\to 0} \frac{1}{t}(f(x,y+t) - f(x,y)). \end{aligned}$$

Especially, we can see that the partial differentiation with respect to a given variable is just the casual one-variable differentiation while considering the other variables to be constants.

8.6. The differential of a function $f : \mathbb{R}^n \to \mathbb{R}$. However, we will not do with partial or directional derivatives for a good approximation of the behavior of a function by linear expressions. We would probably expect that a "differentiable" function of more variables composed with any differentiable curve again yields differentiable functions of one variable, which we have known well.

However, let us look at the functions in the plane given by the formulae

$$g(x,y) = \begin{cases} 1 & \text{if } yx = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h(x,y) = \begin{cases} 1 & \text{if } y = x^2 \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

8.17. 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

$$\begin{aligned} z &= f(2,1) + f_x(2,1)(x - x_0) + f_y(2,1)(y - y_0) \\ &= 5x + 2y - 12. \end{aligned}$$

The tangent plane to the graph of g is then

$$\begin{aligned} z &= f(2,1) + g_x(x,y)(2,1)(x - x_0) + g(x,y)_y(2,1)(y - y_0) \\ &= 2y - 2. \end{aligned}$$

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 $[2, 1 + t, 2t]$, $t \in \mathbb{R}$. □

8.18. 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:

$$f_x = \frac{y\ln(z)}{2\sqrt{xy\ln(z)}},\ f_y = \frac{x\ln(z)}{2\sqrt{xy\ln(z)}},\ f_z = \frac{xy}{2z\sqrt{xy\ln(z)}}.$$

Each of these three partial derivatives is again a function of three variables, so we can consider (first) partial derivatives of these functions. Those are the second partial derivatives of the function f. We will write the variable with respect to which we differentiate as a subscript of the function f.

$$\begin{aligned} f_{xx} &= -\frac{y^2\ln^2 z}{4(xy\ln z)^{\frac{3}{2}}}, \\ f_{xy} &= -\frac{xy\ln^2 z}{4(xy\ln z)^{\frac{3}{2}}} + \frac{\ln z}{2\sqrt{xy\ln z}}, \\ f_{xz} &= -\frac{xy^2\ln z}{4z(xy\ln z)^{\frac{3}{2}}} + \frac{y}{2z\sqrt{xy\ln z}}, \\ f_{yy} &= -\frac{x^2\ln^2 z}{4(xy\ln z)^{\frac{3}{2}}}, \\ f_{yz} &= -\frac{x^2y\ln z}{4z(xy\ln z)^{\frac{3}{2}}} + \frac{x}{2z\sqrt{xy\ln z}}, \end{aligned}$$

Apparently, neither of them extends all smooth curves going through the point $(0, 0)$ to smooth functions. On the other hand, both partial derivatives of g at $(0, 0)$ exist and no other directional derivatives do, while h has all directional derivatives at the point $(0, 0)$, and we even have $d_v h(0) = 0$ for all directions v, so this is a linear dependence on $v \in \mathbb{R}^2$.

We can also imagine a function f which, along the lines $(r\cos\theta, r\sin\theta)$ with a fixed angle θ, takes the values $k(\theta)r$, where $k(\theta)$ 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 will not be linear expressions depending on the directions v for general functions $k(\theta)$.

Therefore, we will imitate the case of univariate functions as thoroughly as possible, and we will forbid such a pathological behavior of functions directly by a definition:

Differential

Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable at a point* x iff 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 increase of v,
(3) $\lim_{v \to 0} \frac{1}{\|v\|}\big(f(x + v) - f(x) - d_v f(x)\big) = 0$.

The linear expression $d_v f$ (in a vector variable v) is called a *differential of the function* f evaluated at the increase of v.

In words, we require that the increases of the function f at the point x be well approximated by linear functions of increases of the variable quantities.

It follows directly from the definition of directional derivatives that the differential can be defined solely by the property (3). Indeed, if there is a linear form $df(x)$ such that the increases 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.

Let us examine what we can say 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 neighborhood of a point (x_0, y_0).

To this purpose, consider any smooth curve $t \mapsto (x(t), y(t))$ with $x_0 = x(0)$, $y_0 = y(0)$. Using the mean value theorem for univariate functions in both summands separately, we obtain that

$$\begin{aligned} &\tfrac{1}{t}\big(f(x(t), y(t)) - f(x_0, y_0)\big) = \\ &\tfrac{1}{t}\big(f(x(t), y(t)) - f(x_0, y(t))\big) + \tfrac{1}{t}\big(f(x_0, y(t)) - f(x_0, y_0)\big) \\ &= \tfrac{1}{t}(x(t) - x_0) \cdot \frac{\partial f}{\partial x}(x(\xi), y(t)) + \tfrac{1}{t}(y(t) - y_0) \cdot \frac{\partial f}{\partial y}(x_0, y(\eta)) \end{aligned}$$

for suitable numbers ξ and η between 0 and t.

Especially, for every sequence of numbers t_n converging to zero, we can get the corresponding sequences of numbers ξ_n and η_n which also converge to zero and all will satisfy the above expression.

Letting $t \to 0$, we get thanks to continuity of the partial derivatives that (see the test for convergence of a function using

$$f_{zz} \;=\; -\frac{x^2y^2}{4z^2(xy\ln z)^{\frac{3}{2}}}-\frac{xy}{2z^2\sqrt{xy\ln z}}.$$

By the theorem about interchangeability of partial derivatives (see 8.10), 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. □

D. Taylor polynomials

8.19. 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

$$\begin{aligned} T_2(x,y) &= f(1,1) + f_x(1,1)(x-1) + f_y(1,1)(y-1) + \\ &\quad + \frac{1}{2}(x-1, y-1)Hf(1,1)\begin{pmatrix} x-1 \\ y-1 \end{pmatrix} \\ &= \ln(3) + \frac{2}{3}(x-1) + \frac{2}{3}(y-1) + \frac{1}{9}(x-1)^2 - \\ &\quad -\frac{4}{9}(x-1)(y-1) + \frac{1}{9}(y-1)^2 \\ &= \frac{1}{9}(x^2+y^2+8x+8y-4xy-14) + \ln(3). \end{aligned}$$

□

Remark. In particular, we can see that the second-order Taylor expansion of an arbitrary differentiable function at a given point is a second-order polynomial.

8.20. 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

$$T(x,y) = \frac{1}{2}\pi^2y^2 - xy - \pi^3y + \frac{1}{2}\pi^4.$$

The tangent plane to the graph of the given function at the point $[\pi, \pi]$ is given by the first-order Taylor polynomial at the point $[\pi, \pi]$; its general equation is thus

$$z = -\pi y - \pi x + \pi^2,$$

and this equation *is* satisfied by the given point $[0, \pi, 0]$. □

subsequences of the input values, 5.23, and Theorem 5.22 about the limits of sums and products of functions)

$$\frac{d}{dt}f(x(t),y(t))|_{t=0} = x'(0)\frac{\partial f}{\partial x}(x_0,y_0) + y'(0)\frac{\partial f}{\partial y}(x_0,y_0),$$

which is a pleasant extension of the theorem on differentiation of composite functions of one variable for vector-valued functions.

Of course, the special choice of parametrized straight lines

$$(x(t), y(t)) = (x_0 + t\xi, y_0 + t\eta)$$

transforms our calculation, with $v = (\xi, \eta)$, to the equality

$$d_vf(x_0,y_0) = \frac{\partial f}{\partial x}(x_0,y_0)\xi + \frac{\partial f}{\partial y}(x_0,y_0)\eta,$$

and this formula can be expressed in the nice way in which we described coordinate expressions of linear functions on vector spaces:

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy.$$

In other words, the directional derivative d_vf is indeed a linear function $\mathbb{R}^n \to \mathbb{R}$ on the increases, with coordinates given by the partial derivatives.

Similarly, we can now prove that the assumption of continuous partial derivatives at a given point guarantees the approximation properties of the differential as well.

We will consider general multivariate functions straightaway:

8.7. Theorem. *Let $f : E_n \to \mathbb{R}$ be a function of n variables which has continuous partial derivatives in a neighborhood of the a point $x \in E_n$. Then its differential df at the point x exists and its coordinate expression is given by the formula*

$$df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_n}dx_n.$$

Proof. This theorem can be derived analogously to the procedure described above, for the case $n = 2$. We only have to be careful in details and finish the reasoning about the approximation properties. Just like above, we consider a curve

$$c(t) = (c_1(t), \dots, c_n(t)),$$

$c(0) = (0, ..., 0)$ and a point $x \in \mathbb{R}^n$, and we express the difference $f(x + c(t)) - f(x)$ for the composite function $f(c(t))$ as follows:

$$\begin{aligned} &f(x_1 + c_1(t), \dots, x_n + c_n(t)) - f(x_1, x_2 + c_2(t), \dots) \\ &\quad + f(x_1, x_2 + c_2(t), \dots)) - f(x_1, x_2, \dots, x_n + c_n(t)) \\ &\quad \vdots \\ &\quad + f(x_1, x_2, \dots, x_n + c_n(t)) - f(x_1, x_2, \dots, x_n). \end{aligned}$$

Now, we can apply the mean value theorem to all of the n summands, thus obtaining (similarly to the case of two variables)

$$\begin{aligned} &(c_1(t) - c_1(0))\frac{\partial f}{\partial x_1}(x_1 + c_1(\theta_1), x_2 + c_2(t), \dots, x_n + c_n(t)) \\ &+ (c_2(t) - c_2(0))\frac{\partial f}{\partial x_2}(x_1, x_2 + c_2(\theta_2), \dots, x_n + c_n(t)) \\ &\vdots \\ &+ (c_n(t) - c_n(0))\frac{\partial f}{\partial x_n}(x_1, x_2, \dots, x_n + c_1(\theta_n)), \end{aligned}$$

8.21. 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 $[0, 0, 0]$. ○

8.22. 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 $[\pi, \pi, \pi]$. ○

8.23. Determine the second-order Taylor polynomial of the function $\ln(x^2 y)$ at the point $[1, 1]$. ○

8.24. 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 $[0, 0]$. ○

E. Extrema of multivariate functions

8.25. 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. □

8.26. 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

$$\begin{aligned} 3y + 10z - 15 &= 0 \\ 2y + 3z - 5 &= 0, \end{aligned}$$

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. □

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.

At the same time, we can apply the mean value theorem in the same way to the difference

$$\begin{aligned} f(x + v) - f(x) &= d_v f(x + \theta v) \\ &= v_1 \frac{\partial f}{\partial x_1}(x + \theta v) + \dots + v_n \frac{\partial f}{\partial x_n}(x + \theta v) \end{aligned}$$

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 thanks to the continuity of the partial derivatives in a neighborhood of the point x.

Since all the partial derivatives are continuous at the point x, we know that for an arbitrarily small $\varepsilon > 0$, there is a neighborhood 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. Thus we get the bound

$$\frac{1}{\|w\|}\big(f(x + w) - f(x) - d_w f(x + \theta w)\big) \le \frac{n}{\|w\|}\|w\|\varepsilon,$$

so the approximation property of the differential is satisfied as well. □

8.8. A plane tangent to the graph of a function. The linear approximation of the function behavior by its differential can, similarly to the case of univariate functions, be expressed with respect to its graph. We will just work with hyperplanes instead of tangent lines.

For 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

$$\begin{aligned} z &= f(x_0, y_0) + df(x_0, y_0)(x - x_0, y - y_0) \\ &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0). \end{aligned}$$

We have 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 out of those which contain the point (x_0, y_0) having the property that all derivatives, and so the tangent lines of all curves

$$c(t) = (x(t), y(t), f(x(t), y(t)))$$

as well lie in it. 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 picture. The diagonal line is the image of the curve $c(t) = (t, t, f(t, t))$.

8.27. 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

$$f_1(x) = x^2 + \arctan^2 x, \quad x \in \mathbb{R}, \qquad f_2(y) = \left| y^3 + y \right|, \quad y \in \mathbb{R}.$$

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

$$f(0, 0) = 0, \qquad f(x, y) > 0, \quad [x, y] \in \mathbb{R}^2 \setminus \{[0, 0]\},$$

the function f has a strict local (even global) minimum at the point $[0, 0]$. □

8.28. Examine the local extrema of the function

$$f(x, y) = \left(x + y^2\right) \mathrm{e}^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

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

$$f_x(x, y) = \mathrm{e}^{\frac{x}{2}} + \tfrac{1}{2}\left(x + y^2\right)\mathrm{e}^{\frac{x}{2}}, \quad f_y(x, y) = 2y\,\mathrm{e}^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

A stationary point $[x, y]$ must satisfy

$$f_y(x, y) = 0, \quad \text{i. e.} \quad y = 0,$$

and, further,

$$f_x(x, y) = f_x(x, 0) = \mathrm{e}^{\frac{x}{2}}\left(1 + \tfrac{1}{2}x\right) = 0, \quad \text{i. e.} \quad x = -2.$$

We can see that there is a unique stationary point, namely $[-2, 0]$.

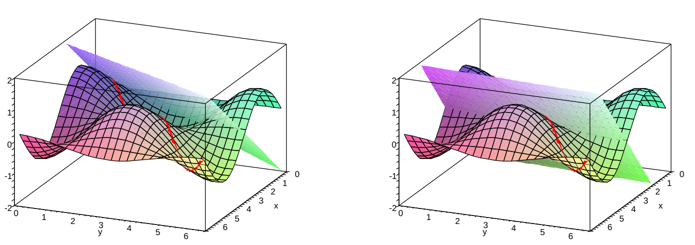
Now, we calculate the Hessian Hf at this point. If this matrix (the corresponding quadratic form) is positive definite, the extremum is a strict local minimum. If it is negative definite, the extremum is a strict local maximum. Finally, if the matrix is indefinite, there will be no extremum at the point. We have

$$f_{xx}(x, y) = \tfrac{1}{2}\mathrm{e}^{\frac{x}{2}}\left(2 + \tfrac{1}{2}\left(x + y^2\right)\right), \quad f_{yy}(x, y) = 2\,\mathrm{e}^{\frac{x}{2}},$$

$$f_{xy}(x, y) = f_{yx}(x, y) = y\,\mathrm{e}^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

Therefore,

$$Hf\,(-2, 0) = \begin{pmatrix} f_{xx}\,(-2, 0) & f_{xy}\,(-2, 0) \\ f_{yx}\,(-2, 0) & f_{yy}\,(-2, 0) \end{pmatrix} = \begin{pmatrix} 1/2\mathrm{e} & 0 \\ 0 & 2/\mathrm{e} \end{pmatrix}.$$



For the case of functions of n variables, the tangent plane is defined as an analogy to the tangent plane to an area in the three-dimensional space. Instead of being puzzled by a great deal of indeces, it will be useful to recall affine geometry, where we already worked with the so-called hyperplanes, see paragraph 4.3.

A TANGENT (HYPER)PLANE TO THE GRAPH OF A FUNCTION AT A POINT

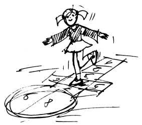
Definition. A *tangent hyperplane* to the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$ is the nadr containing the point $(x, f(x))$ with direction which is the graph of the linear mapping $df(x) : \mathbb{R}^n \to \mathbb{R}$, i. e. the differential at the point $x \in E_n$.

The definition takes advantage of the fact that the directional derivative $d_v f$ is given by the increase in the tangent (hyper)plane corresponding to the increase of the input vector v.

Many analogies with the univariate functions follow from these reasonings. 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 direction 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 like in the case of univariate functions, this depends on the values of higher derivatives.

8.9. Derivatives of higher orders. The operation of differentiation can be iterated similarly to the case of univariate functions. This time, we can choose different directions for each iteration.



If we fix an increase $v \in \mathbb{R}^n$, the enumeration of the differentials at this increase 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, we can repeat this procedure with another increase, and so on. In particular, we can work with iterations of partial derivatives. For *second-order partial derivatives*, we write

$$\left(\frac{\partial}{\partial x_j} \circ \frac{\partial}{\partial x_i}\right) f = \frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

In the case of the repeated choice $i = j$, we also write

$$\left(\frac{\partial}{\partial x_i} \circ \frac{\partial}{\partial x_i}\right) f = \frac{\partial^2}{\partial x_i^2} f = \frac{\partial^2 f}{\partial x_i^2}.$$

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 $[-2, 0]$. □

8.29. 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 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

$$\begin{aligned} 3x^2 - 3z = 0, &\quad \text{i. e.,} \quad z = x^2, \\ 2y - 2 = 0, &\quad \text{i. e.,} \quad y = 1, \end{aligned}$$

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:

$$Hf\,(1, 1, 1) = \begin{pmatrix} 6 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix}, \quad Hf\,(2, 1, 4) = \begin{pmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix}.$$

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

$$d_1 = \left|a_{11}\right|, \; d_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \; d_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix}, \; \dots, \; d_n = |\,A\,|,$$

are positive. Further, it is negative definite iff

$$d_1 < 0, \quad d_2 > 0, \quad d_3 < 0, \quad \dots, \quad (-1)^n d_n > 0.$$

The inequalities

We proceed in the same way with further iterations and talk about *partial derivatives of order k*

$$\frac{\partial^k f}{\partial x_{i_1} \dots \partial x_{i_k}}.$$

More generally, we can also iterate (assuming the function is sufficiently differentiable) any directional derivatives; for instance, $d_v \circ d_w f$ for two fixed increases $v, w \in \mathbb{R}^n$.

k–TIMES DIFFERENTIABLE FUNCTIONS

We say that a function $f : E_n \to \mathbb{R}$ is *k–times (continuously) differentiable* at a point x iff all partial derivatives up to order k (inclusive) exist in a neighborhood of the point x and are continuous at this point.

We say that f is k–differentiable iff it is k–times (continuously) differentiable at all points of its domain.

>From now on, we will always work with continuously differentiable functions unless explicitly stated otherwise.

To show all of this in the simplest form, we will once again 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 indeces 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 will show that if certain senseful conditions are satisfied, the partial derivatives commute, i. e., we need not care about the order in which we differentiate.

Since we suppose that the partial derivatives exist and are continuous, the limits

$$\begin{aligned} f_{xy}(x, y) &= \lim_{t \to 0} \frac{1}{t}\big(f_x(x, y+t) - f_x(x, y)\big) \\ &= \lim_{t \to 0} \frac{1}{t}\Big(\lim_{s \to 0} \frac{1}{s}\big(f(x+s, y+t) - f(x, y+t) \\ &\qquad - f(x+s, y) + f(x, y)\big)\Big) \end{aligned}$$

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, we will also have that

$$\begin{aligned} f_{xy}(x, y) &= \lim_{t \to 0} \frac{1}{t^2}\Big(\big(f(x+t, y+t) - f(x, y+t)\big) \\ &\qquad - \big(f(x+t, y) - f(x, y)\big)\Big), \end{aligned}$$

and this limit value is continuous at (x, y).

Let us consider the expression from which we take the last limit to be a function $\varphi(x, y, t)$, and let us try to express it in terms of partial derivatives. For a temporarily fixed t, we denote $g(x, y) = f(x+t, y) - f(x, y)$. Then the expression in the last big 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).

$$\left|12\right| = 12 > 0, \quad \begin{vmatrix} 12 & 0 \\ 0 & 2 \end{vmatrix} = 24 > 0, \quad \begin{vmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{vmatrix} = 6 > 0,$$

imply that the matrix $Hf\ (2,1,4)$ is positive definite – there is a strict local minimum at the point $[2,1,4]$. □

8.30. 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:

$$Hz\ (0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad Hz\ (1,0) = Hz\ (-1,0) = \begin{pmatrix} -16 & 0 \\ 0 & 0 \end{pmatrix}.$$

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

$$z\ (1,0) = z\ (-1,0) = 0, \qquad z\ (x,0) < 0 \quad \text{for } x \in (-1,1).$$

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

$$z\ \left(x, \sqrt{2\left(1 - x^4\right)}\right) = \left(x^2 - 1\right)\left(x^4 - 1\right) > 0, \quad x \in (-1,1).$$

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. □

8.31. 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. □

8.32. Determine local extrema of the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2y + y^2z + x - z$ on $\mathbb{R}^3$. ○

Now, $g_y(x, y) = f_y(x + t, y) - f_y(x, y)$, so we can write φ as

$$\begin{aligned} \varphi(x, y, t) &= \frac{1}{t} g_y(x, y + t_0) \\ &= \frac{1}{t}\big(f_y(x + t, y + t_0) - f_y(x, y + t_0)\big). \end{aligned}$$

Another application of the mean value theorem yields

$$\varphi(x, y, t) = f_{yx}(x + t_1, y + t_0)$$

for a suitable t_1 between 0 and t. However, if we split the big parentheses 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 we assume that the second-order partial derivatives are continuous, the limit for $t \to 0$ must guarantee the wanted equality

$$f_{xy}(x, y) = f_{yx}(x, y)$$

at all points (x, y).

The same procedure for functions of n variables proves the following fundamental result:

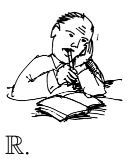
Interchangeability of partial derivatives

8.10. 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 neighborhood 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 was illustrated above in the special case when $n = 2$. The procedure works similarly for the general case as well.

Formally, the proof can be led in the following way: we may assume that for every fixed choice of a pair of coordinates x_i and x_j, the whole discussion of their interchanging takes place in a two-dimensional affine subspace, i. e., all the other variables are considered to be constant and take no effect in the reasonings.

In the case of higher-order derivatives, the proof can be finished by induction on the order. Indeed, every order of the indeces $i_1, \dots, i_k$ can be obtained from a fixed one by several swaps of adjacent pairs of indeces. □



8.11. Hessian. In the case of first-order derivatives, we introduced the differential, being the linear form $df(x)$ which approximates a function f at a point x in the best way. Similarly, we will now want to understand the quadratic approximation of a function $f : E_n \to \mathbb{R}$.

8.33. 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 $\mathbb{R}^3$. ○

8.34. 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 $\mathbb{R}^3$. ○

8.35. 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 $\mathbb{R}^3$. ○

8.36. 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 $\mathbb{R}^2$ ○

8.37. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2y + 2y^2 + 2y$ on $\mathbb{R}^2$. ○

8.38. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + xy + 2y^2 + y$ on $\mathbb{R}^2$. ○

8.39. Determine the local extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + xy - 2y^2 + y$ on $\mathbb{R}^2$. ○

F. Implicitly given functions and mappings

8.40. 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 $[1, 1]$:

$$F_y(x, y) = x\sin\left(\frac{\pi}{2}xy^2\right) + \pi x^2y^2\cos\left(\frac{\pi}{2}xy^2\right),$$

so $F_y(1, 1) = 1 \neq 0$. Therefore, it follows from theorem 8.18 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.

8.41. 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:

$$f'(0) = -\frac{F_x(0, \pi/2)}{F_y(0, \pi/2)} = -\frac{1}{1} = -1. \qquad \square$$

Hessian

Definition. If $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function, we call the symmetric matrix of functions

$$Hf(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1\partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_1\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n\partial x_n}(x) \end{pmatrix}$$

the Hessian of the function f at the point x.

We have already seen from the previous reasonings that zeroing the differential at a point $(x, y) \in E_2$ guarantees stationary behavior along all curves going through this point. The Hessian

$$Hf(x, y) = \begin{pmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{xy}(x, y) & f_{yy}(x, y) \end{pmatrix}$$

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 univariate functions

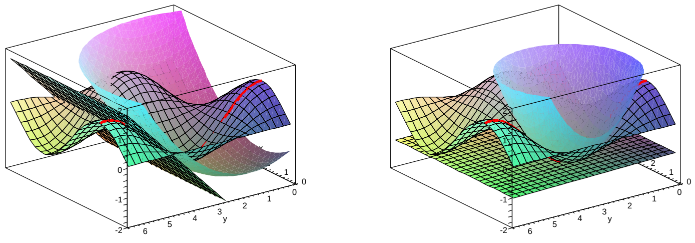
$$\begin{aligned} \alpha(t) &= f(x(t), y(t)) \\ \beta(t) &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)\xi + \frac{\partial f}{\partial y}(x_0, y_0)\eta \\ &\quad + \frac{1}{2}\Big(f_{xx}(x_0, y_0)\xi^2 + 2f_{xy}(x_0, y_0)\xi\eta + f_{yy}(x_0, y_0)\eta^2\Big) \end{aligned}$$

will share 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

$$\beta(t) = f(x_0, y_0) + df(x_0, y_0)\cdot\begin{pmatrix}\xi\\ \eta\end{pmatrix} + \frac{1}{2}(\xi\ \eta)\cdot Hf(x_0, y_0)\cdot\begin{pmatrix}\xi\\ \eta\end{pmatrix}$$

or $\beta(t) = f(x_0, y_0) + df(x_0, y_0)(v) + \frac{1}{2}Hf(x_0, y_0)(v, v)$, where $v = (\xi, \eta)$ is the increase 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 picture shows both the tangent plane and this quadratic approximation for two distinct points and the function $f(x, y) = \sin(x)\cos(y)$.



8.12. Taylor's theorem. The multidimensional version of Taylor's theorem is once again an example of a mathematical statement where the most difficult part is finding the right formulation. The proof is quite simple then.

8.42. 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 $[\pi, 1, 0]$.

$$\begin{aligned} F_x(x, y, z) &= y\cos(xy) + z\cos(xz) \;\; F_x(\pi, 1, 0) = -1, \\ F_y(x, y, z) &= x\cos(xy) + z\cos(yz) \;\; F_y(\pi, 1, 0) = -\pi, \end{aligned}$$

odkud

$$\begin{aligned} z_x(\pi, 1) &= -\frac{F_x(\pi, 1, 0)}{F_z(\pi, 1, 0)} = \frac{1}{\pi + 1}, \\ z_y(\pi, 1) &= -\frac{F_y(\pi, 1, 0)}{F_z(\pi, 1, 0)} = \frac{\pi}{\pi + 1}. \end{aligned}$$

□

8.43. 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 8.18) that the equation $F(x, c_1(x), c_2(x)) = (0, 0)$ on a neighborhood of the point $[1, \pi, 1]$ determines a curve $(c_1(x), c_2(x))$ defined on a neighborhood of the point $[1, \pi]$. In order to find its tangent vector at this point, we need to determine the)column) vector (f_x, g_x) at this point:

$$\begin{pmatrix} f_x \\ g_x \end{pmatrix} = \begin{pmatrix} \sin y\, e^{x\sin y} \\ yz \end{pmatrix}, \quad \begin{pmatrix} f_x(1, \pi, 1) \\ g_x(1, \pi, 1) \end{pmatrix} = \begin{pmatrix} 0 \\ \pi \end{pmatrix}.$$

The wanted tangent vector is thus

$$\begin{aligned} \begin{pmatrix} (c_1)_x(1) \\ (c_2)_x(1)) \end{pmatrix} &= \begin{pmatrix} f_y(1, \pi, 1) & f_z(1, \pi, 1) \\ g_y(1, \pi, 1) & g_z(1, \pi, 1) \end{pmatrix}^{-1} \begin{pmatrix} f_x(1, \pi, 1 \\ g_x(1, \pi, 1) \end{pmatrix} = \\ &= \begin{pmatrix} -1 & 0 \\ 1 & \pi \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \pi \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ \frac{1}{\pi} & \frac{1}{\pi} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{aligned}$$

□

We will proceed in the direction mentioned above, and we will introduce a notation for the particular parts of $D^k f$ approximations of higher orders for functions $f : E_n \to \mathbb{R}^n$. It will alwyas be k–linear expressions in the increases, and we will be interested only in their enumeration at a k-tuple of same values.

We have already discussed the differential $D^1 f = df$ (the first order) and the Hessian $D^2 f = Hf$ (the second order). In general, for functions $f : E_n \to \mathbb{R}$, points $x = (x_1, \dots, x_2) \in E_n$, and increases $v = (\xi_1, \dots, \xi_n)$, we set

$$D^k f(x)(v) = \sum_{1 \le i_1, \dots, i_k \le n} \frac{\partial^k f}{\partial x_{i_1} \dots \partial x_{i_k}}(x_1, \dots, x_n) \cdot \xi_{i_1} \cdots \xi_{i_k}.$$

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 neighborhood $\mathcal{O}_\delta(x)$ of a point $x \in E_n$. For every increase $v \in \mathbb{R}^n$ of size $\|v\| < \delta$, there exists a number θ, $0 \le \theta \le 1$, such that*

$$\begin{aligned} f(x + v) = f(x) + D^1 f(x)(v) &+ \frac{1}{2!} D^2 f(x)(v) + \\ \dots + \frac{1}{(k-1)!} D^{k-1} f(x)(v) &+ \frac{1}{k!} D^k f(x + \theta \cdot v)(v). \end{aligned}$$

Proof. For an increase $v \in \mathbb{R}^n$, we consider the parametrized straight line $c(t) = x + tv$ in E_n, and we 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 6.4)

$$\begin{aligned} \varphi(t) = \varphi(0) &+ \varphi'(0)t + \dots \\ &+ \frac{1}{(k-1)!}\varphi^{(k-1)}(0)t^{k-1} + \frac{1}{k!}\varphi^{(k)}(\theta)t^k. \end{aligned}$$

Therefore, it remains to verify that step-by-step differentiation of the composite function φ yields just the wanted 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 we have already used several times. When deriving it, we used the formula

$$\frac{d}{dt}\varphi(t) = \frac{\partial f}{\partial x_1}(x(t)) \cdot x_1'(t) + \dots + \frac{\partial f}{\partial x_n}(x(t)) \cdot x_n'(t),$$

G. Constrained optimization

We will begin with a somewhat atypical optimization problem.

8.44. A betting office accepts bets on the outcome of a tennis match. Let the odds laid against player A winning be $a : 1$ (i. e., if a bettor bets x dollars on the event that player A wins and this really happens, then the bettor wins ax dollars) and, similarly, let the odds laid against player B winning be $b : 1$ (fees are neglected). What is the necessary and sufficient condition for (positive real) numbers a and b so that a bettor cannot guarantee any profit regardless the actual outcome of the match? (For instance, if the odds were laid $1.5 : 1$ against the win of A and $5 : 1$ against the win of B, then the bettor could bet 3 dollars on B winning and 7 dollars on A winning and profit from this bet in either case).

Solution. Let the bettor have P dollars. The bet amount can be divided to kP and $(1-k)P$ dollars, where $k \in (0, 1)$. The profit is then akP dollars (if player A wins) or $b(1-k)P$ dollars (if B does). The bettor is always guaranteed to win the lesser of these two amounts; the total profit (or loss) is obtained by subtracting the bet P, then. Since each of a, b, P is a positive real number, the function akP is increasing, and the function $b(1-k)P$ is decreasing with respect to k. For $k = 0$, $b(1-k)P$ is greater; for $k = 1$, akP is. The minimum of the two numbers akP and $b(1-k)P$ is thus maximal for a $k \in (0, 1)$, namely for the value k_0 which satisfies $ak_0P = b(1-k_0)P$, whence $k_0 = \frac{b}{a+b}$. Therefore, the betting office must choose a, b so that $ak_0P = b(1-k_0)P < P$, which is equivalent to $ak_0 < 1$, i. e., $ab < a + b$. □

We managed to solve this constrained optimization problem even without using the differential calculus. However, we will not be able to do so in the following problems.

8.45. 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

$$\begin{aligned}0 &= 3x^2 - 2\lambda x,\\ 0 &= 3y^2 - 2\lambda y,\\ 0 &= 3z^2 - 2\lambda z,\\ 0 &= x^2 + y^2 + z^2 - 1\end{aligned}$$

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 S. Therefore, we are likely to get 6 local

which holds for every continuously differentiable curve and function f. This means that

$$D^1 f(c(t))(v) = D^1 f(c(t))(c'(t))$$

for all t in a neighborhood of zero. We will proceed accordingly for functions $D^\ell f$. We can write $c'(t)$ instead of the increase v, and we will remember 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).

We suppose that

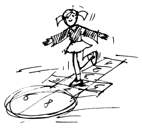
$$D^\ell f(x)(v) = \sum_{1\le i_1,\dots,i_\ell\le n}\Big(\frac{\partial^\ell f}{\partial x_{i_1}\dots\partial x_{i_\ell}}(x_1(t),\dots,x_n(t)) \cdot x'_{i_1}(t)\cdots x'_{i_\ell}(t)\Big),$$

let us calculate this for $\ell + 1$. By the formula for first-order differentiation in a given direction, which has been derived, and the rule for the derivative of a product (see Theorem 5.33), differentiation of the composite function gives

$$\begin{aligned}&\frac{d}{dt}D^\ell f(c(t))(c'(t)) =\\ &= \frac{d}{dt}\sum_{1\le i_1,\dots,i_\ell\le n}\Big(\frac{\partial^\ell f}{\partial x_{i_1}\dots\partial x_{i_\ell}}(x_1(t),\dots,x_n(t)) \cdot x'_{i_1}(t)\cdots x'_{i_\ell}(t)\Big)\\ &= \sum_{1\le i_1,\dots,i_\ell\le n}\Big(\sum_{j=1}^{n}\frac{\partial^{\ell+1} f}{\partial x_{i_1}\dots\partial x_{i_\ell}\partial x_j}(x_1(t),\dots,x_n(t)) \cdot x'_j(t)\cdot x'_{i_1}(t)\cdots x'_{i_\ell}(t)\Big) + 0,\end{aligned}$$

which is indeed the wanted formula for the 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. □

8.13. Local extrema of multivalued functions. Now, we will try to examine the local maxima and minima of functions on E_n using the differential and the Hessian. Just like in the case of univariate functions, we say that an interior point $x_0 \in E_n$ of the domain of a function f is a (local) *maximum* or *minimum* iff there is a neighborhood U of its such that for all points $x \in U$, the function value satisfies $f(x) \le f(x_0)$ or $f(x) \ge f(x_0)$, respectively. If equality holds for no $x \ne x_0$ in the previous inequalities, we talk about a *strict extremum*.



For the sake of simplicity, we will suppose that our function f has continuous both first-order and second-order partial derivatives on its domain. A necessary condition for existence of an extremum at a point x_0 is that the differential be zero at this point, i. e., $df(x_0) = 0$. Indeed, if $df(x_0) \ne 0$, then there is a direction v in which we have $d_v f(x_0) \ne 0$. 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 (5.32).

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*.

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

$$\begin{aligned}
0 &= 3x^2 - 2\lambda x - \eta,\\
0 &= 3y^2 - 2\lambda y - \eta,\\
0 &= 3z^2 - 2\lambda z \ - \eta,\\
0 &= x^2 + y^2 + z^2 - 1,\\
0 &= x + y + z.
\end{aligned}$$

However, since a circle is also a compact set, h must have both a global minimum and maximum on it. Further analysis is left to the reader. □

8.46. 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:

$$\begin{aligned}
2xy &= 4kx,\\
x^2 &= 4ky,\\
0 &= 2kz.
\end{aligned}$$

This system is satisfied by the points $[\pm\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{6}}, 0]$ and $[\pm\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{6}}, 0]$. The function takes on only two values at these four stationary points. 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. □

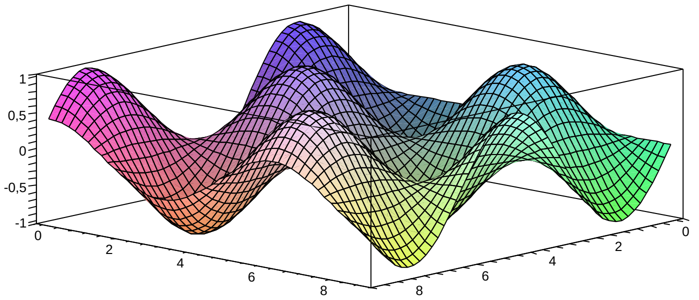
8.47. Decide whether the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = z - xy^2$ has any minima and maxima on the sphere

$$x^2 + y^2 + z^2 = 1.$$

If so, determine them.

We will again, for a while, work with a simple function in E_2 in order to illustrate our conclusions directly. Let us consider the function $f(x, y) = \sin(x)\cos(y)$ which has been discussed and caught in many pictures, namely in paragraphs 8.9 a 8.8.

The shape of this function resembles well-known egg plates, so it is apparent that we can find a lot of extrema, but also many more stationary point which, in fact, will not be extrema (the little "saddles" noticeable in the picture).



Therefore, let is 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 will be 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

$$\begin{aligned}
Hf(x, y) &= \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}(x, y)\\
&= \begin{pmatrix} -\sin(x)\cos(y) & -\cos(x)\sin(y) \\ -\cos(x)\sin(y) & -\sin(x)\cos(y) \end{pmatrix}.
\end{aligned}$$

We thus get the following Hessians in our two sets of stationary points:

(1) $Hf(k\pi + \frac{\pi}{2}, \ell\pi) = \pm\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, where the sign $-$ occurs when k and ℓ have the same parity (remainder upon 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 sign $-$ occurs occurs when k and ℓ have the same parity, and the sign $+$ occurs in the other case;

Now, if we look at the proposition of Taylor's theorem for order $k = 2$, we get, in a neighborhood of one of the stationary points (x_0, y_0),

$$\begin{aligned}
&f(x, y) = f(x_0, y_0)+\\
&+\frac{1}{2}Hf(x_0 + \theta(x - x_0), y_0 + \theta(y - y_0))(x - x_0, y - y_0),
\end{aligned}$$

where Hf is now considered a quadratic form evaluated at the increase $(x - x_0, y - y_0)$. Since the Hessian of our function is continuous (i. e., continuous partial derivatives up to order two, inclusive) and the matrices of the Hessian are non-degenerate, the local maximum occurs if and only if our point (x_0, y_0) belongs to the former group with k and ℓ of the same parity. On the other

Solution. We are looking for solutions of the system

$$\begin{aligned} x &= -ky^2, \\ y &= -2kxy, \\ z &= k. \end{aligned}$$

The second equation implies that either $y = 0$ or $x = -\frac{1}{2k}$. The first possibility leads to the points $[0, 0, 1]$, $[0, 0, -1]$. The second one cannot be satisfied (substituting into the equation of the sphere, we get the equation

$$\frac{1}{4k^2} + \frac{1}{2k^2} + k^2 = 1,$$

which has no solution. The function has a maximum and minimum, respectively, at the two computed points on the given sphere. □

8.48. 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:

$$\begin{aligned} \frac{\partial g}{\partial x} = \lambda \frac{\partial f}{\partial x} : \quad yz &= 2\lambda k x, \\ \frac{\partial g}{\partial y} = \lambda \frac{\partial f}{\partial y} : \quad xz &= 2\lambda l y, \\ \frac{\partial g}{\partial z} = \lambda \frac{\partial f}{\partial z} : \quad xy &= 2\lambda z. \end{aligned}$$

We can easily see that the equation can only be satisfied by a triple of non-zero numbers. Dividing pairs of equations and substituting into the ellipse's equation, we get eight solutions, namely the stationary points $x = \pm\frac{1}{\sqrt{3k}}$, $y = \pm\frac{1}{\sqrt{3l}}$, $z = \pm\frac{1}{\sqrt{3}}$. However, the function f takes on only two distinct values at these eight points. Since it is continuous and the given ellipsoid is compact, f must have both a maximum and minimum on it. Moreover, since both f and g are continuously differentiable, these extrema must occur at stationary points. Therefore, it must be that four of the computed stationary points are local maxima of the function (of value $\frac{1}{3\sqrt{3kl}}$) and the other four are minima (of value $-\frac{1}{3\sqrt{3kl}}$). □

8.49. 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

$$x \geq 0, \quad y \geq 0, \quad y \leq -x + 3. \tag{8.1}$$

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 (‖8.1‖) is clearly a triangle with vertices at $[0, 0]$, $[3, 0]$, $[0, 3]$.

hand, if the parities are different, then the point from the former group happens to be a point of a local minimum.

On the other hand, the Hessian of the latter group of points is always positive at some increases and negative at other ones. Therefore, the entire function f behaves in this manner in a small neighborhood of the given point.

In order to formulate the general statement about the Hessian and the local extrema at stationary points, we have to remember the discussion about quadratic forms from the paragraphs 4.31–4.32 in the chapter on affine geometry. There, we introduced the following attributes for a quadratic form $h : E_n \to \mathbb{R}$:

- *positively definite* iff $h(u) > 0$ for all $u \neq 0$
- *positively semidefinite* iff $h(u) \geq 0$ for all $u \in V$
- *negatively definite* iff $h(u) < 0$ for all $u \neq 0$
- *negatively semidefinite* iff $h(u) \leq 0$ for all $u \in V$
- *indefinite* iff $h(u) > 0$ and $f(v) < 0$ for appropriate $u, v \in V$.

We also invented some methods which allow us to find out whether a given form has any of these properties.

The Taylor expansion with remainder immediately yields the following proposition:

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 positively definite,

(2) f has a strict local minimum at x if $Hf(x)$ is negatively 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 out function $f(x_1, \dots, x_n)$, an arbitrary point $x = (x_1, \dots, x_n)$, and any increase $v = (v_1, \dots, v_n)$, such that both x and $x + v$ 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, $0 \leq \theta \leq 1$. Since we suppose that the differential is zero, we get

$$f(x + v) = f(x) + \frac{1}{2}Hf(x + \theta \cdot v)(v).$$

By our 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 4.32. However, the determinant itself is a polynomial expression in the coefficients of the matrix, hence a continuous function. Therefore, the non-zeroness and signs of the examined determinants are the same in a sufficiently small neighborhood of the point x as at the point x itself.

In particular, for positively definite $Hf(x)$, we have 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. □

Let us determine the stationary points inside this triangle as the solution of the equations $f_x = 0$, $f_y = 0$. Since

$$f_x(x, y) = 2x + 4y - 6, \quad f_y(x, y) = 4x - 4y,$$

these equations are satisfied only by the point $[1, 1]$. The boundary suggests itself to be expressed as the union of three line segments given by the choice of pairs of vertices. First, we consider $x = 0$, $y \in [0, 3]$, when $f(x, y) = -2y^2 - 1$. However, we know the graph of this (univariate) function on the interval $[0, 3]$ It is thus not difficult to find the points at which global extrema occur. They are the marginal points $[0, 0]$, $[0, 3]$. Similarly, we can consider $y = 0$, $x \in [0, 3]$, also obtaining the marginal points $[0, 0]$, $[3, 0]$. Finally, we get to the line segment $y = -x + 3$, $x \in [0, 3]$. Making some rearrangements, we get

$$f(x, y) = f(x, -x + 3) = -5x^2 + 18x - 19, \quad x \in [0, 3].$$

We thus need to find the stationary points of the polynomial $p(x) = -5x^2 + 18x - 19$ from the interval $[0, 3]$. The equation $p'(x) = 0$, i. e., $-10x + 18 = 0$, is satisfied by $x = 9/5$. This means that in the last case, we obtained one more point (besides the marginal points), namely $[9/5, 6/5]$, where a global extremum may occur. Altogether, we have these points as "suspects":

$$[1, 1], \quad [0, 0], \quad [0, 3], \quad [3, 0], \quad \left[\tfrac{9}{5}, \tfrac{6}{5}\right]$$

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 $[0, 3]$. □

8.50. 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:

$$\begin{aligned} 0 &= 2k + l, \\ 2yz &= k - l, \\ y^2 &= k + 2l, \\ 2x + y + z &= 1, \\ x - y + 2z &= 0. \end{aligned}$$

The solution is $[x, y, z] = [\frac{2}{3}, 0, -\frac{1}{3}]$ and $[x, y, z] = [\frac{4}{9}, \frac{2}{9}, -\frac{1}{9}]$ (of course, the variables k and l can also be computed, but we are not interested in them). The marginal points of the given line segment are $[-1, \frac{5}{3}, \frac{4}{3}]$ and $[2, -\frac{4}{3}, -\frac{5}{3}]$. Considering these four points, the function takes on the greatest value at the first marginal point ($f(x, y, z) = \frac{100}{27}$), which is its maximum on the given segment, and it

Let us notice that the theorem yields no result if the Hessian of the examined function is degenerate, yet not indefinite at the point in question. The reason is again 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, we cannot determine whether the function behaves like t^3 or $\pm t^4$ until we calculate the higher-order derivatives in the necessary directions at least.

At the same time, we can notice that even at those points where the differential is non-zero, the definiteness of the Hessian $Hf(x)$ has similar consequences as the non-zeroness of the second derivative of a univariate function. Indeed, for a function $f : \mathbb{R}^n \to \mathbb{R}$, the expression

$$z(x + v) = f(x) + df(x)(v)$$

defines a tangent hyperplane to the graph of the function f in the space $\mathbb{R}^{n+1}$, so Taylor's theorem of order two with remainder, as used in the proof, shows that when the Hessian is positively definite, all the values of the function f lie in a sufficiently small neighborhood of the point x above the values of the tangent hyperplane, i. e., the whole graph is above the tangent hyperplane in a sufficiently small neighborhood. In the case of negative definiteness, it is the other way round. Finally, when the Hessian is indefinite, the graph of the function goes from one side of the hyperplane to the other, but this happens, in general, along objects of lower dimension in the tangent hyperplane, so we have no straightforward generalization of inflexion points.

8.14. The differential of mappings. The concepts of a derivative and a 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

$$F(x_1, \dots, x_n) = (f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n))$$

of functions $f_i : E_n \to \mathbb{R}$. We say that F is a *differentiable* or *k–times differentiable mapping* iff the corresponding property is shared by all the functions $f_1, \dots, f_m$.

Differential and Jacobian matrix

The differentials $df_i(x)$ of the particular functions f_i give a linear approximation of the increases of their values for the mapping

$$F(x_1, \dots, x_n) = (f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n)).$$

Therefore, we can expect that they will also give a coordinate expression of the linear mapping $D^1F(x) : \mathbb{R}^n \to \mathbb{R}^m$ between the direction spaces, which linearly approximates the increases of our mapping. The resulting 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 *Jacobian matrix of the mapping F* at a point x.

The linear mapping $D^1F(x)$ defined on the increases $v = (v_1, \dots, v_n)$ by identically denoted the Jacobian matrix is called takes the least value at the second marginal point ($f(x, y, z) = -\frac{80}{27}$), which is thus its minimum there. □

8.51. Find the maximal and minimal values of the polynomial

$$p(x, y) = 4x^3 - 3x - 4y^3 + 9y$$

on the set

$$M = \left\{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \le 1\right\}.$$

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

$$\tfrac{1}{\cos^2 t} = 1 + \tan^2 t,$$

which is valid provided both sides are well-defined. We thus obtain

$$f'(t) =$$
$$\cos^3 t\left[-12\tan t + 3\left(\tan t + \tan^3 t\right) - 12\tan^2 t + 9\left(1 + \tan^2 t\right)\right]$$

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

$$-4\tan t + \tan t + \tan^3 t - 4\tan^2 t + 3 + 3\tan^2 t = 0.$$

The substitution $s = \tan t$ 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

$$s = 1, \quad s = \sqrt{3}, \quad s = -\sqrt{3}$$

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

$$f\left(-\tfrac{1}{3}\pi\right) = -1 - 3\sqrt{3} < f\left(-\tfrac{3}{4}\pi\right) = -3\sqrt{2} < f\left(-\tfrac{2}{3}\pi\right) = 1 - 3\sqrt{3} < -1,$$

$$f(-\pi) = f(\pi) = -1 < 0,$$

$$f\left(\tfrac{2}{3}\pi\right) = 1 + 3\sqrt{3} > f\left(\tfrac{1}{4}\pi\right) = 3\sqrt{2} > f\left(\tfrac{1}{3}\pi\right) = -1 + 3\sqrt{3} > 0.$$

the *differential of the mapping* F at a point x in the domain iff we have

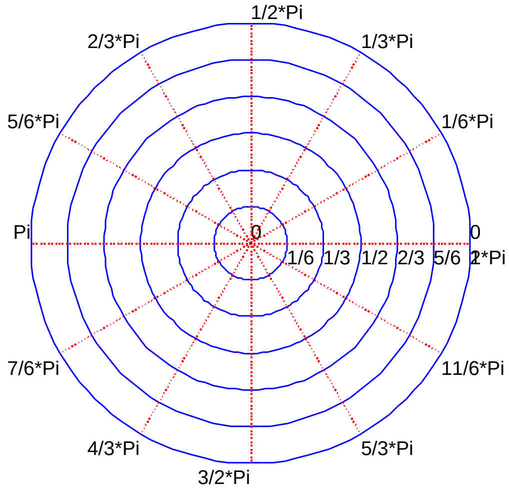
$$\lim_{v \to 0} \frac{1}{\|v\|}\left(F(x+v) - F(x) - D^1F(x)(v)\right) = 0.$$

Several times, we have already used the fact that the definition of Euclidean distance guarantees that the limits of values in E_n exist if and only if the limits of the particular coordinate components do. Direct application of Theorem 8.5 about existence of the differential for functions of n variables to the particular coordinate functions of the mapping F thus leads to the following proposition (prove this in detail by yourselves!):

Corollary. *Let $F : E_n \to E_m$ be a mapping such that all of its coordinate functions have continuous partial derivatives in a neighborhood of a point $x \in E_n$. Then the differential $D^1F(x)$ exists, and it is given by the Jacobian matrix $D^1F(x)$.*

8.15. 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 whole of F's image is called a *transformation*. Such a mapping can be perceived as a change of coordinates. We usually require that both F and G be (continuously) differentiable mappings.

Just like in the case of vector spaces, the choice of our "point of view", i. e. the choice of coordinates, can simplify or deteriorate our 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. A very illustrative example is the change of the most usual coordinates in the plane to the so-called polar ones, i. e., 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 change from 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

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 $\cos\left(-\frac{1}{3}\pi\right) = \frac{1}{2}$, $\sin\left(-\frac{1}{3}\pi\right) = -\frac{\sqrt{3}}{2}$, $\cos\left(\frac{2}{3}\pi\right) = -\frac{1}{2}$, $\sin\left(\frac{2}{3}\pi\right) = \frac{\sqrt{3}}{2}$, we can deduce that the polynomial p takes on the minimal value $-1 - 3\sqrt{3}$ (the same as f, of course) at the point $[1/2, -\sqrt{3}/2]$ and the maximal value $1 + 3\sqrt{3}$ at $[-1/2, \sqrt{3}/2]$. □

8.52. 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 $[-1, 0]$. □

8.53. Compute the local extrema of the function $y = f(x)$ given implicitly by the equation

$$3x^2 + 2xy + x = y^2 + 3y + \tfrac{5}{4}, \quad [x, y] \in \mathbb{R}^2 \smallsetminus \left\{\left[x, x - \tfrac{3}{2}\right];\ x \in \mathbb{R}\right\}.$$

Solution. In accordance with the theoretical part (see 8.18), let us denote

$$F(x, y) = 3x^2 + 2xy + x - y^2 - 3y - \tfrac{5}{4},$$
$$[x, y] \in \mathbb{R}^2 \smallsetminus \left\{\left[x, x - \tfrac{3}{2}\right];\ x \in \mathbb{R}\right\}$$

and calculate the derivative

$$y' = f'(x) = -\tfrac{F_x(x,y)}{F_y(x,y)} = -\tfrac{6x+2y+1}{2x-2y-3}.$$

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

$$y'' = \left(y'\right)' = -\tfrac{(6+2y')(2x-2y-3)-(6x+2y+1)(2-2y')}{(2x-2y-3)^2}.$$

Substituting $x = 0$, $y = -1/2$, $y' = 0$ and $x = 1/2$, $y = -2$, $y' = 0$, we obtain

$$y'' = -\tfrac{6(-2)-0}{4} > 0 \quad \text{for } [x, y] = \left[0, -\tfrac{1}{2}\right]$$

and

$$y'' = -\tfrac{6(+2)-0}{4} < 0 \quad \text{for } [x, y] = \left[\tfrac{1}{2}, -2\right].$$

mapping would exist. The Cartesian image of lines in polar coordinates with constant coordinates r or φ is shown in the picture above.

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 a bit complicated, it is actually the same as the one we have already seen in the case of one variable. The Jacobian matrix for univariate functions is in fact a single number, namely the derivative of the function at a given point, so the multiplication of Jacobian 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 formulae we have derived for the derivative of a composition of multivariate functions with a curve.

The differential of a composite mapping

8.16. 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 form the domain of F is given by the composition of differentials*

$$D^1(G \circ F)(x) = D^1G(F(x)) \circ D^1F(x).$$

The corresponding Jacobian matrix is given by the product of the corresponding Jacobian matrices.

Proof. In paragraph 8.5 and in the proof of Taylor's theorem, we derived how differentiation of mappings composed from functions and curves behaves This proves the special cases of this theorem for $n = r = 1$. The general case can be proved analogously, we just have to work with more vectors now.

Let us fix an arbitrary increase v and calculate the directional derivative for the composition $G \circ F$ at a point $x \in E_n$. This actually means to determine the differentials for the particular coordinate functions of the mapping G composed with F. For the sake of simplicity, we will 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

$$\begin{aligned} g(F(x + tv)) - g(F(x) &= dg(F(x))(F(x + tv) - F(x)) \\ &\quad + \alpha(F(x + tv) - F(x)), \end{aligned}$$

where α is a function defined on a neighborhood of the point $F(x)$ which is continuous and satisfies $\lim_{v \to 0} \frac{1}{\|v\|}\alpha(v) = 0$. Substitution into the equality for the directional derivative yields

$$\begin{aligned} d_v(g \circ F)(x) &= \lim_{t \to 0} \frac{1}{t}\big(dg(F(x))(F(x + tv) - F(x)) \\ &\qquad + \alpha(F(x + tv) - F(x))\big) \\ &= dg(F(x))\left(\lim_{t \to 0} \frac{1}{t}\Big(F(x + tv) - F(x)\Big)\right) \\ &\qquad + \lim_{t \to 0} \frac{1}{t}\Big(\alpha(F(x + tv) - F(x))\Big) \\ &= dg(F(x)) \circ D^1F(x)(v) + 0, \end{aligned}$$

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 $x = 1/2$. □

8.54. Find the local extrema of the function $z = f(x, y)$ given on the maximum possible set by the equation

$$x^2 + y^2 + z^2 - xz - yz + 2x + 2y + 2z - 2 = 0. \tag{8.2}$$

Solution. Differentiating (‖8.2‖) 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

$$\begin{aligned} z_x = f_x(x, y) &= \frac{z \ - 2x - 2}{2z - x - y + 2}, \\ z_y = f_y(x, y) &= \frac{z \ - 2y - 2}{2z - x - y + 2}. \end{aligned} \tag{8.3}$$

We can notice that the partial derivatives are continuous at all points where the function f is defined. This implies that the local extrema can occur only at stationary points. These points satisfy

$$z_x = 0, \quad \text{i. e.} \quad z \ - 2x - 2 = 0,$$
$$z_y = 0, \quad \text{i. e.} \quad z \ - 2y - 2 = 0.$$

We have thus two equations, which allow us to express the dependency of x and y on z. Substituting into (‖8.2‖), 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 (‖8.3‖), we obtain

$$z_{xx} = f_{xx}(x, y) = \tfrac{(z_x - 2)(2z - x - y + 2) - (z - 2x - 2)(2z_x - 1)}{(2z - x - y + 2)^2},$$

with respect to x, and

$$z_{xy} = f_{xy}(x, y) = \tfrac{z_y(2z - x - y + 2) - (z - 2x - 2)(2z_y - 1)}{(2z - x - y + 2)^2},$$

with respect to y. We need not calculate z_{yy} since the variables x and y are interchangeabel in (‖8.2‖) (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:

$$f_{xx}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = f_{yy}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = -\tfrac{1}{\sqrt{6}},$$
$$f_{xy}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = f_{yx}\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = 0,$$
$$f_{xx}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = f_{yy}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = \tfrac{1}{\sqrt{6}},$$
$$f_{xy}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = f_{yx}\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = 0.$$

As for the Hessian, we have

$$Hf\left(-3 + \sqrt{6}, -3 + \sqrt{6}\right) = \begin{pmatrix} -\frac{1}{\sqrt{6}} & 0 \\ 0 & -\frac{1}{\sqrt{6}} \end{pmatrix},$$
$$Hf\left(-3 - \sqrt{6}, -3 - \sqrt{6}\right) = \begin{pmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

where we made use of the properties of the function α and the fact the a linear mapping between finite-dimensional spaces are always continuous.

Thus, we have proved the theorem for the particular functions $g_1, \dots, g_r$ of the mapping G. The whole theorem now follows from matrix multiplication. □

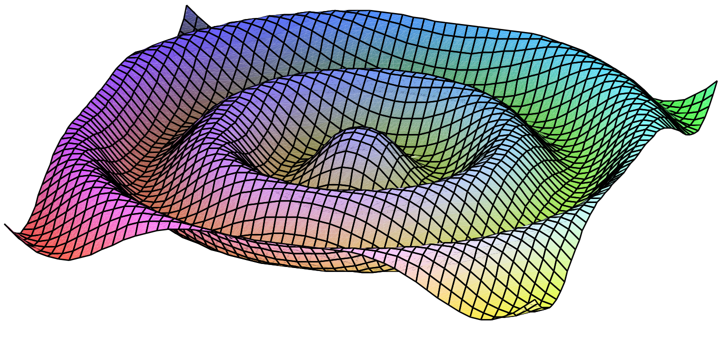
Now, we can illustrate, by a simple example, the usage of our concept of transformation and the theorem about differentiation of composite mappings. We have seen that the polar coordinates are given from the Cartesian ones by the transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$ which, in coordinates (x, y) and (r, φ), is written as follows (for instance, on the domain of all point in the first quadrant except for the points having $x = 0$):

$$r = \sqrt{x^2 + y^2}, \ \varphi = \arctan\frac{y}{x}.$$

Consider a function $g_t : E_2 \to \mathbb{R}$ which can be expressed as

$$g(r, \varphi, t) = \sin(r - t)$$

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 picture (there, $t = -\pi/2$). While it was easy to define it in polar coordinates, it would have been much harder in Cartesian ones.



Now, let us compute the derivative of this function in Cartesian coordinates. Using our theorem, we get

$$\begin{aligned} \frac{\partial g}{\partial x}(x, y, t) &= \frac{\partial g}{\partial r}(r, \varphi)\frac{\partial r}{\partial x}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi)\frac{\partial \varphi}{\partial x}(x, y) \\ &= \cos(\sqrt{x^2 + y^2} - t)\frac{x}{\sqrt{x^2 + y^2}} + 0 \end{aligned}$$

and, similarly,

$$\begin{aligned} \frac{\partial g}{\partial y}(x, y, t) &= \frac{\partial g}{\partial r}(r, \varphi)\frac{\partial r}{\partial y}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi)\frac{\partial \varphi}{\partial y}(x, y) \\ &= \cos(\sqrt{x^2 + y^2} - t)\frac{y}{\sqrt{x^2 + y^2}}. \end{aligned}$$

8.17. The inverse mapping theorem. If the first derivative of a univariate function is non-zero, its sign determines whether the function is increasing or decreasing. Then, the function must have this property in a neighborhood of the point in question, and so an inverse function exists in the selected neighborhood. The derivative of the

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 $\left[-3-\sqrt{6}, -3-\sqrt{6}\right]$. □

8.55. Determine the strict local 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. Since both the function f and the function given implicitly by the equation $\frac{1}{x^2} + \frac{1}{y^2} - 4 = 0$ have continuous partial derivatives of all orders on the set $\mathbb{R}^2 \setminus \{[0,0]\}$, we should look for stationary points, i. e., for the solution of the equations $L_x = 0$, $L_y = 0$ for

$$L(x,y,\lambda) = \tfrac{1}{x} + \tfrac{1}{y} - \lambda\left(\tfrac{1}{x^2} + \tfrac{1}{y^2} - 4\right), \quad x \neq 0, \ y \neq 0.$$

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

$$\left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right], \quad \left[-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right]. \tag{8.4}$$

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

$$d^2L(x,y) = \left(\tfrac{2}{x^3} - \tfrac{6\lambda}{x^4}\right)dx^2 + \left(\tfrac{2}{y^3} - \tfrac{6\lambda}{y^4}\right)dy^2.$$

Differentiating the constraint $\frac{1}{x^2} + \frac{1}{y^2} = 4$, 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,

$$d^2L(x,y) = \left[\tfrac{2}{x^3} - \tfrac{6\lambda}{x^4} + \left(\tfrac{2}{y^3} - \tfrac{6\lambda}{y^4}\right)\tfrac{y^6}{x^6}\right]dx^2.$$

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

$$d^2L\left(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2}\right) = -4\sqrt{2}\,dx^2, \quad d^2L\left(-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2}\right) = 4\sqrt{2}\,dx^2,$$

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:

$$f\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = 2\sqrt{2}, \quad f\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) = -2\sqrt{2}. \tag{8.5}$$

Now, we will demonstrate a quicker way how to obtain the result. We know (or we can easily calculate) the second partial derivatives of the function L, i. e., the Hessian with respect to the variables x and y:

$$HL(x,y) = \begin{pmatrix} \frac{2}{x^3} - \frac{6\lambda}{x^4} & 0 \\ 0 & \frac{2}{y^3} - \frac{6\lambda}{y^4} \end{pmatrix}.$$

inverse function is then the multiplicative inverse of the derivative of the original function.

Interpreting this situation for a mapping $E_1 \to E_1$ and linear mappings $\mathbb{R} \to \mathbb{R}$ as their differentials, the non-zeroness is a necessary and sufficient condition for the differential to be invertible. In this way, we obtain a statement which is valid for finite-dimensional spaces in general:

The inverse mapping theorem

Theorem. *Let $F : E_n \to E_n$ be a differentiable mapping on a neighborhood of a point $x_0 \in E_n$, and let the Jacobian matrix $D^1 f(x_0)$ be invertible. Then in some neighborhood of the point x_0, the inverse mapping F^{-1} exists, and its differential at the point $F(x_0)$ is the inverse mapping to the differential $D^1F(x_0)$, i. e., it is given by the inverse matrix to the Jacobian matrix of the mapping F at the point x_0.*

Proof. First, we should try to verify that the theorem makes sense and is expectable. If we supposed that the inverse mapping existed and was differentiable at the point $F(x_0)$, differentiation of composite functions 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 end of the theorem. Therefore, we know right from the beginning which differential for F^{-1} to look for.

In the next step, we will suppose that the inverse mapping F^{-1} exists in a neighborhood of the point $F(x_0)$ and that it is continuous. We are to verify the existence of the differential. Since F is differentiable in a neighborhood of x_0, it follows that

$$F(x) - F(x_0) - D^1F(x_0)(x - x_0) = \alpha(x - x_0)$$

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 $y_0 = F(x_0)$:

$$\lim_{y \to y_0} \frac{1}{\|y - y_0\|}\left(F^{-1}(y) - F^{-1}(y_0) - (D^1F(x_0))^{-1}(y - y_0)\right).$$

Substitution into the previous equality gives

$$\begin{aligned} &\lim_{y \to y_0} \frac{1}{\|y - y_0\|}\Big(x - x_0 - \\ &\qquad (D^1F(x_0))^{-1}(D^1F(x_0)(x - x_0) + \alpha(x - x_0))\Big) \\ &= \lim_{y \to y_0} \frac{-1}{\|y - y_0\|}(D^1F(x_0))^{-1}(\alpha(x - x_0)) \\ &= (D^1F(x_0))^{-1} \lim_{y \to y_0} \frac{-1}{\|y - y_0\|}(\alpha(x - x_0)), \end{aligned}$$

where the last equality follows from the fact that linear mappings between finite-dimensional spaces are always continuous, and thanks to invertibility of the differential, performing it before the limit process has no impact upon existence of the limit.

The evaluation

$$HL\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)=\begin{pmatrix}-2\sqrt{2} & 0\\ 0 & -2\sqrt{2}\end{pmatrix},$$
$$HL\left(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)=\begin{pmatrix}2\sqrt{2} & 0\\ 0 & 2\sqrt{2}\end{pmatrix}$$

then tells us that the quadratic form is negative definite for the former stationary point (there is a strict local maximum) and positive definite for the latter one (there is a strict local minimum).

We should be aware of a potential trap in this "quicker" method in the case we obtain an indefinite form (matrix). Then, we cannot conclude that there is not an extremum at that point since as we have not included the constraint (which we did when computing d^2L), we are considering a more general situation. The graph of the function f on the given set is a curve which can be defined as a univariate function. This must correspond to a one-dimensional quadratic form. □

8.56. 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 (‖8.4‖) and the values (‖8.5‖) 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 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

$$f_{\max}=f\left(\tfrac{\sqrt{2}}{2},\tfrac{\sqrt{2}}{2}\right)=2\sqrt{2},\quad f_{\min}=f\left(-\tfrac{\sqrt{2}}{2},-\tfrac{\sqrt{2}}{2}\right)=-2\sqrt{2}.$$

Let us notice that we are almost done with the proof. The limit at the end of our expression is, thanks to the properties of the function α, zero if the magnitudes $\|F(x)-F(x_0)\|$ are greater than $C\|x-x_0\|$ for some constant C. This is a bit stronger property than F^{-1} being continuous; in literature, this property of a function is called *Lipschitz continuity*. So, now it remains "merely" to prove the existence of a Lipschitz-continuous inverse mapping to the mapping F.

To simplify the reasonings to come, we will transform the general case to a statement which is a bit more simple. Especially, we can achieve $x_0=0\in R^n$, $y_0=F(x_0)=0\in\mathbb{R}^n$ by a convenient choice of Cartesian coordinates, which is without loss of generality.

Composing the mapping F with any linear mapping G yields a differentiable mapping again, and we know this changes the differential. The choice $G(x)=(D^1F(0))^{-1}(x)$ gives $D^1(G\circ F)(0)=\mathrm{id}_{\mathbb{R}^n}$. Therefore, we can assume that

$$D^1F(0)=\mathrm{id}_{\mathbb{R}^n}\,.$$

Now, having these assumptions, let us consider the mapping $K(x)=F(x)-x$. This mapping is differentiable, too, and its differential at 0 is apparently zero.

By The Taylor expansion with remainder of the particular coordinate functions K_i and the definition of Euclidean distance, we get for any continuously differentiable mapping K in a neighborhood of the origin of $\mathbb{R}^n$ the bound

$$\|K(x)-K(y)\|\leq C\sqrt{n}\|x-y\|,$$

where C is bounded by the maximum of all absolute values of the partial derivatives in the Jacobian matrix of the mapping K in the neighborhood in question.[2]

Since the differential of the mapping K at the point $x_0=0$ is zero in our case, we can, selecting a sufficiently small neighborhood U of the origin, achieve the bound

$$\|K(x)-K(y)\|\leq\frac{1}{2}\|x-y\|.$$

Further, substituting for the definition $K(x)=F(x)-x$ and invoking the triangle inequality $\|(u-v)+v\|\leq\|u-v\|+\|v\|$, i. e., $\|u\|-\|v\|\leq\|u-v\|$ as well, we get

$$\begin{aligned}\|y-x\|-\|F(x)-F(y)\|&\leq\|F(x)-F(y)+y-x\|\\&\leq\frac{1}{2}\|y-x\|.\end{aligned}$$

Hence, finally,

$$\frac{1}{2}\|x-y\|\leq\|F(x)-F(y)\|.$$

With this bound, we have reached a great advancement: if $x\neq y$ in our neighborhood U of the origin, then we also must have $F(x)\neq F(y)$. Therefore, our mapping is bijective. Let us write F^{-1} for its inverse defined on the image of U. For this function, our bound says that

$$\|F^{-1}(x)-F^{-1}(y)\|\leq 2\|x-y\|,$$

so this mapping is not only continuous, but also Lipschitz-continuous, as we needed in the previous part of the proof.

[2] It immediately follows from this reasoning that a function which has continuous partial derivatives on a compact set is Lipschitz-continuous on it as well.

□

8.57. 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

$$L(x, y, z, \lambda_1, \lambda_2) = xyz - \lambda_1 \left(x^2 + y^2 + z^2 - 1\right) - \lambda_2 (x + y + z),$$

the equations $L_x = 0$, $L_y = 0$, $L_z = 0$ give

$$\begin{aligned} yz - 2\lambda_1 x - \lambda_2 &= 0, \\ xz - 2\lambda_1 y - \lambda_2 &= 0, \\ xy - 2\lambda_1 z - \lambda_2 &= 0, \end{aligned}$$

respectively. Subtracting the first equation from the second one and from the third one leads to

$$\begin{aligned} xz - yz - 2\lambda_1 y + 2\lambda_1 x &= 0, \\ xy - yz - 2\lambda_1 z + 2\lambda_1 x &= 0, \end{aligned}$$

i. e.,

$$\begin{aligned} (x - y)\,(z + 2\lambda_1) &= 0, \\ (x - z)\,(y + 2\lambda_1) &= 0. \end{aligned}$$

The last equations are satisfied in these four cases:

$$\begin{aligned} x = y, x = z; &\quad x = y, y = -2\lambda_1; \\ z = -2\lambda_1, x = z; &\quad z = -2\lambda_1, y = -2\lambda_1, \end{aligned}$$

thus (including the constraint $G = 0$)

$$\begin{aligned} x = y = z = 0; &\quad x = y = -2\lambda_1, z = 4\lambda_1; \\ x = z = -2\lambda_1, y = 4\lambda_1; &\quad x = 4\lambda_1, y = z = -2\lambda_1. \end{aligned}$$

Except for the first case (which clearly cannot happen), including the constraint $F = 0$ yields

$$(4\lambda_1)^2 + (-2\lambda_1)^2 + (-2\lambda_1)^2 = 1, \quad \text{i. e.} \quad \lambda_1 = \pm\tfrac{1}{2\sqrt{6}}.$$

Altogether, we get the points

$$\left[-\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}\right], \quad \left[-\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right], \quad \left[\tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right],$$
$$\left[\tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}\right], \quad \left[\tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right], \quad \left[-\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right].$$

It could seem that we are done with the proof, yet it is not true. To finish the proof completely, we still have to show that the mapping F restricted to a sufficiently small neighborhood is not only bijective, but also that it maps open neighborhoods of zero onto open neighborhoods of zero. [3]

Let us choose δ so small that the neighborhood $V = \mathcal{O}_\delta(0)$ lies in U with its boundary, and, at the same time, the Jacobian matrix of the mapping is invertible on the whole V. This surely can be done since the determinant is a continuous mapping. Let B denote the boundary of the set V (i. e., the corresponding sphere). Since B is compact and F is continuous, the function

$$\rho(x) = \|F(x)\|$$

has both a maximum and a minimum on B. Let us denote $a = \frac{1}{2}\min_{x\in B}\rho(x)$ and consider any $y \in \mathcal{O}_a(0)$. Of course, $a > 0$. We want to show that there is at least one $x \in V$ such that $y = F(x)$, which will prove the whole inverse mapping theorem.

To this purpose, consider the function (y is our fixed point)

$$h(x) = \|F(x) - y\|^2.$$

Again, the image $h(V) \cup h(B)$ must have a minimum. First, we show that this minimum cannot occur for $x \in B$. Indeed, we have $F(0) = 0$, hence $h(0) = \|y\| < a$. At the same time, by our definition of a, the distance of y from $F(x)$ for $x \in B$ is at least a for $y \in \mathcal{O}_a(0)$ (since a was selected to be half the minimum of the magnitude of $F(x)$ on the boundary). Therefore, the minimum occurs inside V, and it must be at a stationary point z of the function h. However, this means that for all $j = 1, \dots, n$, we have

$$\frac{\partial h}{\partial x^j}(z) = \sum_{i=1}^{n} 2(f_i(z) - y_i)\frac{\partial f_i}{\partial x_j}(z) = 0.$$

This system of equations can be considered a system of linear equations with variables $\xi_i = f_i(z) - y_i$ and coefficients given by twice the Jacobian matrix $D^1F(z)$. For every $z \in V$, such a system has a unique solution, and that is zero since we suppose that the Jacobian matrix is invertible.

Thus, we have found the wanted point $x = z \in V$ satisfying, for all $i = 1, \dots, n$, the equality $f_i(z) = y_i$, i. e., $F(z) = y$. □

8.18. The implicit function theorem. Our next goal is to apply the inverse mapping theorem for work with implicitly defined functions. For the beginning, let us consider a differentiable function $F(x, y)$ defined in the plane E_2, and let us look for those point (x, y) at which $F(x, y) = 0$.

An example of this can be the usual (implicit) definition of straight lines and circles:

$$\begin{aligned} F(x, y) &= ax + by + c = 0 \\ F(x, y) &= (x - s)^2 + (y - t)^2 - r^2 = 0, \ r > 0. \end{aligned}$$

While in the first case, the function given by the first formula is (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

[3] In literature, there are many examples of mappings which, for instance, continuously and bijectively map a line segment onto a square.

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

$$f\left(\frac{1}{\sqrt{6}},\frac{1}{\sqrt{6}},-\frac{2}{\sqrt{6}}\right)=f\left(\frac{1}{\sqrt{6}},-\frac{2}{\sqrt{6}},\frac{1}{\sqrt{6}}\right)= \\ =f\left(-\frac{2}{\sqrt{6}},\frac{1}{\sqrt{6}},\frac{1}{\sqrt{6}}\right)=-\frac{1}{3\sqrt{6}}.$$

□

8.58. 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. □

8.59. 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

$$\left[\frac{1}{3}\pm\frac{\sqrt{22}}{6},\frac{1}{3}\pm\frac{\sqrt{22}}{6},\frac{1}{3}\mp\frac{\sqrt{22}}{3}\right].$$

points of the circle in the direction of the coordinate x), we can find a neighborhood 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 which semicircle the point (x_0,y_0) belongs to. Having the picture of the situation drawn, the reason is clear: we cannot describe both the semicircles simultaneously by a single function $y=f(x)$. The marginal points of the interval $[s-r,s+r]$ are more amazing. They also satisfy the equation of the circle, yet we have at them that $F_y(s\pm r,t)=0$, which describes the position of the tangent line to the circle at these points, parallel to the y-axis. Indeed, we cannot find neighborhoods of these points in which the circle could be described as a function $y=f(x)$.

Moreover, the derivatives of our function $y=f(x)=t+\sqrt{(x-s)^2-r^2}$ at points where it is defined can be expressed in terms of partial derivatives of the function F:

$$f'(x)=\frac{1}{2}\frac{2(x-s)}{\sqrt{(x-s)^2-r^2}}=\frac{x-s}{y-t}=-\frac{F_x}{F_y}.$$

If we interchange the roles of the variables x and y and we will want to find a dependency $x=f(y)$ such that $F(f(y),y)=0$, then we will succeed in neighborhoods of points $(s\pm r,t)$ with no problem. Let us notice that the partial derivative F_x is non-zero at these points.

Our observation thus (for mere two examples) says: for a function $F(x,y)$ and a point $(a,b)\in E_2$ such that $F(a,b)=0$, there is a unique function $y=f(x)$ satisfying $F(x,f(x))=0$ if we have $F_y(a,b)\neq 0$. In this case, we can even compute $f'(a)=-F_x(a,b)/F_y(a,b)$. We will prove that actually, this proposition is always true. The last statement about derivatives can be remembered (and is quite comprehensible if things are thoroughly understood) from the expression for the differential of the function $g(x)=F(x,y(x))$ and the differential $dy=f'(x)dx$

$$0=dg=F_xdx+F_ydy=(F_x+F_yf'(x))dx.$$

We could work analogously with the implicit expressions $F(x,y,z)=0$, where we can look for a function $g(x,y)$ such that $F(x,y,g(x,y))=0$. As an example, consider the function $f(x,y)=x^2+y^2$, whose graph is a circular 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 straight for the general situation, we can notice which dimensions could/should appear in the problem. If we wanted 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 we succeed as well (even for all initial conditions $x=a$), yet the result will not be unique for a given initial condition. In fact, it suffices to consider an arbitrary curve on the circular paraboloid whose projection onto the first coordinate has non-zero derivative. Then we consider x to be the parameter of the curve, and $c(x)$ is chosen to be its projection onto the plane yz.

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. □

8.60. 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. ○

8.61. 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 $[x_0, y_0] = [0, 1]$. ○

8.62. 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 $[x_0, y_0] = [2, 0]$. ○

8.63. 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 $[x_0, y_0] = [2, 0]$. ○

8.64. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = xy$ on the ellipse $x^2 + 3y^2 = 1$. ○

8.65. Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f = x^2y$ on the circle with radius 1 which is centered at the point $[x_0, y_0] = [0, 0]$. ○

8.66. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^3y$ on the circle $x^2 + y^2 = 1$. ○

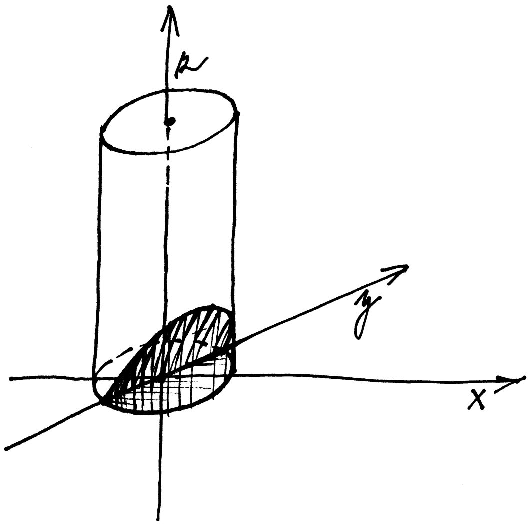
8.67. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = xy$ on th ellipse $2x^2 + 3y^2 = 1$. ○

8.68. Find the maximum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = xy$ on the ellipse $x^2 + 2y^2 = 1$. ○

H. Volumes, areas, centroids of solids

8.69. 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) $x + y + z \leq 0$.



Therefore, we expect that one function of $m + 1$ variables defines implicitly a hypersurface in $\mathbb{R}^{m+1}$ which we want to express (at least locally) as the graph of a function of m variables. We can anticipate that n functions of $m + n$ variables will define an intersection of n hypersurfaces in $\mathbb{R}^{m+n}$, which is, in "most" cases, a m–dimensional object.

Let us thus consider a differentiable mapping

$$F = (f_1, \dots, f_n) : \mathbb{R}^{m+n} \to \mathbb{R}^n.$$

The Jacobian matrix of this mapping will have n rows and $m + n$ columns, and we can write it symbolically as

$$D^1F = (D^1_xF, D^1_yF)$$
$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_m} & \frac{\partial f_1}{\partial x_{m+1}} & \dots & \frac{\partial f_1}{\partial x_{m+n}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_m} & \frac{\partial f_n}{\partial x_{m+1}} & \dots & \frac{\partial f_n}{\partial x_{m+n}} \end{pmatrix},$$

where $(x_1, \dots, x_{m+n}) \in \mathbb{R}^{m+n}$ is written as $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$, D^1_xF is a matrix of n rows and the first m columns in the Jacobian matrix, while D^1_yF 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^1_yF is invertible.

The implicit mapping theorem

Theorem. *Let $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$ be a differentiable mapping in an open neighborhood of a point $(a, b) \in \mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$ at which $F(a, b) = 0$, and* $\det D^1_yF \neq 0$. *Then there exists a differentiable mapping $G : \mathbb{R}^m \to \mathbb{R}^n$ defined on an neighborhood 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 Jacobian matrix D^1G of the mapping G is, in the neighborhood of the point a, given by the product of matrices

$$D^1G(x) = -(D^1_yF)^{-1}(x, G(x)) \cdot D^1_xF(x, G(x)).$$

Proof. For the sake of comprehensibility, we 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 will be quite complicated because it will be presented in a way which can be extended for the general dimensions as the theorem states.

We extend the function F to

$$\tilde{F} : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x, y) \mapsto (x, F(x, y)).$$

The Jacobian 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 holds in a neighborhood of the point (a, b) as well, so the function $\tilde{F}$ is invertible in this neighborhood, by the inverse mapping theorem. Therefore, let us take the uniquely defined differentiable inverse mapping $\tilde{F}^{-1}$ in a neighborhood of the point $(a, 0)$.

Now, let us denote by $\pi : \mathbb{R}^2 \to \mathbb{R}$ the projection onto the second coordinate, and consider the function $f(x) = \pi \circ \tilde{F}^{-1}(x, 0)$.

Solution. a) The volume can be calculated with ease using cylindric coordinates. There, the cylinder is determined by the inequality $r \leq 1$; the half-plane $z \leq x$ by $z \leq r\cos\varphi$, then. Altogether, we get

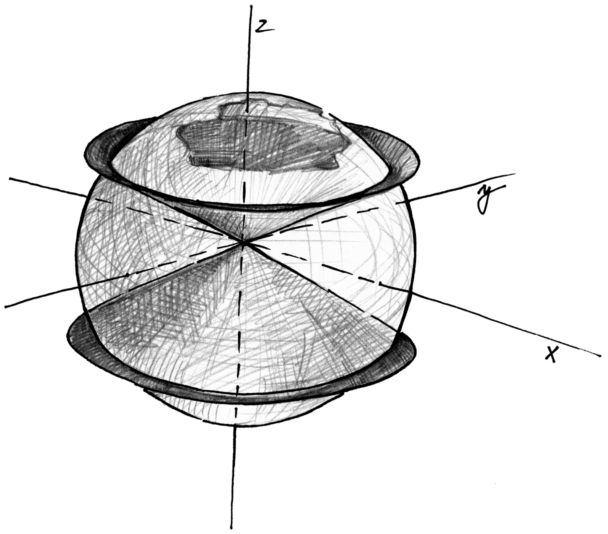
$$V = \int_0^1 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{r\cos\varphi} r \,\mathrm{d}z \,\mathrm{d}\varphi \,\mathrm{d}r = \frac{2}{3}.$$

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 ||8.79||. □

8.70. Find the volume of the solid in $\mathbb{R}^3$ which is given by $x^2 + y^2 + z^2 \leq 1$, $3x^2 + 3y^2 \geq z^2$, $x \geq 0$.

Solution.



First, we should realize what the examined solid looks like. It is a part of a ball which lies outside a given cone (see the picture).

The best way to determine the volume is probably to subtract half the volume of the sector given by the cone from half the ball's volume (note that the volume of the solid does not change if we replace the condition $x \geq 0$ with $z \geq 0$ – the sector is cut either "horizontally" or "vertically", but always to halves). We will calculate in spherical coordinates.

$$\begin{aligned} x &= r\cos(\varphi)\sin(\psi), \\ y &= r\sin(\varphi)\sin(\psi), \\ z &= r\cos(\psi), \end{aligned}$$

$\varphi \in [0, 2\pi)$, $\psi \in [0, \pi)$, $r \in (0, \infty)$.

The Jacobian of this transformation $\mathbb{R}^3 \to \mathbb{R}^3$ is $r^2 \sin(\psi)$.

This function is well-defined and differentiable. We must verify that the expression

$$F(x, f(x)) = F(x, \pi(\tilde{F}^{-1}(x, 0)))$$

evaluates to zero in a neighborhood 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, we can resume the previous calculation:

$$\begin{aligned} F(x, f(x)) = \pi(\tilde{F}(x, \pi(\tilde{F}^{-1}(x, 0)))) &= \\ &= \pi(\tilde{F}(\tilde{F}^{-1}(x, 0))) = \pi(x, 0) = 0. \end{aligned}$$

This proves the first part of the theorem, and it remains to compute the derivative of the function $f(x)$. This derivative can, once again, be obtained by invoking the inverse mapping theorem, using the matrix $(D^1\tilde{F})^{-1}$.

The following result can be 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 2.23)

$$\begin{pmatrix} 1 & 0 \\ F_x(x, y) & F_y(x, y) \end{pmatrix}^{-1} = (F_y(x, y))^{-1} \begin{pmatrix} F_y(x, y) & 0 \\ -F_x(x, y) & 1 \end{pmatrix}.$$

By the definition $f(x) = \pi\tilde{F}^{-1}(x, 0)$, we are interested in the first entry of the second row of this matrix, which is just the Jacobian matrix $D^1 f$. In our simple case, it is exactly the wanted 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 mentioned formulae (the participating objects just get the "vector" sense), except for the last computation of the derivative of the function. There, the corresponding parts of the Jacobian matrix $D_x^1 F$ and $D_y^1 F$ will appear, instead of the particular partial derivatives Of course, we need to work with vectors and matrices, rather than scalars.

For the calculation of the Jacobian matrix of the mapping G, we will once again use the computation of an inverse matrix, yet the procedure from paragraph 2.23 is not very advantageous. It is much better to get inspired 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 (i. e., A, for instance, is of type $m \times m$, while C is of type $n \times m$). Now, we can determine the matrices A, B, C, D from the definition equality for an 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 from here that $A = \mathrm{id}_{\mathbb{R}^m}$, $B = 0$, $D = (D_y^1 F)^{-1}$, and, finally, $D_x^1 F + D_y^1 F \cdot C = 0$. From the last equality, we get the desired relation

$$D^1 G = C = -(D_y^1 F)^{-1} \cdot D_x^1 F.$$

This proves the theorem. □

First of all, let us determine the volume of the ball. As for the integration bounds, it is convenient to express the conditions that bind the solid in the coordinates we will work in. In the spherical coordinates, the ball is given by the inequality

$$x^2 + y^2 + z^2 = r^2 \leq 1.$$

First, let us find the integration bounds for the variable 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 $\pi_\varphi(\mathit{ball}) = [0, 2\pi)$ (the equation $r^2 \leq 1$ 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 $r^2 \leq 1$ or the equality $\varphi = \varphi_0$), so it can take on all possible values, $\psi \in [0, \pi)$.

Finally, let us fix a $\varphi = \varphi_0$ and a $\psi = \psi_0$. Now, we are looking for the projection $\pi_r(U)$ of the object (line segment) U given by the constraints $r^2 \leq 1$, $\varphi = \varphi_0$, $\psi = \psi_0$ on the variable r. The only constraint for r is the condition $r^2 \leq 1$, so $r \in (0, 1]$.

Note that the integration bounds of the variables are independent of each other, so we can perform the integration in any order. Thus, we have

$$V_{\text{koule}} = \int_0^1 \int_0^{2\pi} \int_0^{\pi} r^2 \sin(\psi)\, \mathrm{d}\psi\, \mathrm{d}\varphi\, \mathrm{d}r = \frac{4}{3}\pi.$$

Now, let us compute the volume of the spherical sector given by $x^2 + y^2 + z^2 \leq 1$ and $3x^2 + 3y^2 \geq z^2$. Again, we express the conditions in the spherical coordinates: $r^2 \leq 1$, $3\sin^2(\psi) \geq \cos^2(\psi)$, i. e., $\tan(\psi) \geq \frac{1}{\sqrt{3}}$. 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 $\tan(\psi) \geq \frac{1}{\sqrt{3}}$, we have $\psi \in [0, \frac{\pi}{6}]$. The variable φ is not restricted by any condition, so $\varphi \in [0, 2\pi]$.

$$V_{\text{sector}} = \int_0^{2\pi} \int_0^1 \int_0^{\frac{\pi}{6}} r^2 \sin\psi\, \mathrm{d}\psi\, \mathrm{d}r\, \mathrm{d}\varphi = \frac{2 - \sqrt{3}}{3}\pi,$$

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:

$$V = \int_0^{\pi} \int_0^1 \int_{\frac{\pi}{6}}^{\frac{5\pi}{6}} r^2 \sin\psi\, \mathrm{d}\psi\, \mathrm{d}r\, \mathrm{d}\varphi = \frac{\pi}{\sqrt{3}}.$$

8.19. The gradient of a function. As we have seen in the previous paragraph, if F is a continuously differentiable function of n variables, the definition $F(x_1, \dots, x_n) = b$ with a fixed value $b \in \mathbb{R}$ defines a subset $M \subset \mathbb{R}^n$ which often has the properties of an $(n-1)$–dimensional hypersurface. To be more precise, if the vector of the partial derivatives

$$D^1F = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$$

is non-zero, we can describe the set M locally as the graph of a continuously differentiable function of $n-1$ variables. In this connection, we also talk about *level sets* M_b. The vector $D^1F \in \mathbb{R}^n$ is called the *gradient of the function* F. In technical and physical literature, it is also often denoted as grad F.

Since M_b is given by a constant value of the function F, the derivatives of the curves lying in M will surely have the property that the differential dF always evaluates to zero on them. Indeed, for every such a curve, we have $F(c(t)) = b$, hence

$$\frac{d}{dt}F(c(t)) = dF(c'(t)) = 0.$$

On the other hand, we can consider a general vector $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ and the magnitude of the corresponding directional derivative

$$|d_vF| = \left|\frac{\partial f}{\partial x_1}v_1 + \dots + \frac{\partial f}{\partial x_n}v_n\right| = \cos\varphi \|D^1F\| \|v\|,$$

where φ is the angle of the vector v from the gradient F, see the discussion about angles of vectors and straight lines in the fourth chapter (definition 4.18).

Hence it follows that the partial derivatives that are zero are exactly those which are perpendicular to the gradient, while the direction given by the gradient is the direction in which the function f increases most rapidly.

Therefore, it is clear that the tangent plane to a non-empty level set M_b in a neighborhood of its point with non-zero gradient D^1F is determined by the orthogonal complement to the gradient, and the gradient itself is the so-called *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 by the equation

$$F(x, y, z) = (x-a)^2 + (y-b)^2 + (z-c)^2 = r^2,$$

we get the normal vectors at a point $P = (x_0, y_0, z_0)$ as a non-zero multiple of the gradient, i. e., a multiple of

$$D^1F = (2(x_0 - a), 2(y_0 - b), 2(z_0 - c)),$$

and the tangent vectors will be 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:

In cylindric coordinates

$$\begin{aligned} x &= r\cos(\varphi),\\ y &= r\sin(\varphi),\\ z &= z \end{aligned}$$

with Jacobian r of this transformation, the calculation of the volume as the difference of the two solids considered above looks as follows:

$$V = \frac{2}{3}\pi - \int_0^{2\pi}\int_0^{\frac{1}{2}}\int_0^1 r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{\pi}{\sqrt{3}}.$$

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 $r \le \frac{1}{2}$ and $r \ge \frac{1}{2}$, respectively.

$$\begin{aligned} V &= V_1 + V_2\\ &= \int_0^{2\pi}\int_0^{\frac{1}{2}}\int_0^{\sqrt{3}r} r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi + \int_0^{2\pi}\int_{\frac{1}{2}}^{1}\int_0^{\sqrt{1-r^2}} r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi\\ &= \frac{\pi}{\sqrt{3}}. \end{aligned}$$

□

Another alternative is to compute it as the volume of a solid of revolution, again splitting the solid into the two parts as in the previous case (the part "under the cone" and the part "under the sphere". However, these solids cannot be obtained by rotating around one of the axes. The volume of the former part can be calculated as the difference between the volumes of the cylinder $x^2 + y^2 \le \frac{1}{4}, 0 \le z \le \frac{\sqrt{3}}{2}$ and the cone's part $3x^2 + 3y^2 \le z^2, 0 \le z \le \frac{\sqrt{3}}{2}$. 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 $y = \sqrt{(1-x^2)}, \frac{1}{2} \le x \le 1$ around the z-axis and the cylinder $x^2 + y^2 \le \frac{1}{4}, 0 \le z \le \frac{\sqrt{3}}{2}$.

$$\begin{aligned} V &= V_1 + V_2\\ &= \left(\frac{\pi\sqrt{3}}{8} - \frac{\pi\sqrt{3}}{24}\right) + \left(\pi\int_0^{\frac{\sqrt{3}}{2}}(1-r^2)\,\mathrm{d}r - \frac{\pi\sqrt{3}}{8}\right)\\ &= \frac{\pi\sqrt{3}}{4} + \frac{\pi}{4\sqrt{3}} = \frac{\pi}{\sqrt{3}}. \end{aligned}$$

8.71. Calculate the volume of the spherical segment of the ball $x^2 + y^2 + z^2 = 2$ cut by the plane $z = 1$.

Tangent hyperplane to an implicitly given hypersurface

Theorem. *For a function $F(x_1,\dots,x_n)$ of n variables and a point $P = (a_1,\dots,a_n)$ in a level set M_b of the function F such that M_b is the graph of a function of $(n-1)$ variables in a neighborhood of the point P, the implicit equation for the tangent hypersurface to M_b is*

$$0 = \frac{\partial f}{\partial x_1}(P)\cdot(x_1 - a_1) + \dots + \frac{\partial f}{\partial x_n}(P)\cdot(x_n - a_n).$$

Proof. The statement is apparent from the previous reasonings. 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 grad F). Clearly, The selected point P satisfies our equation. □

8.20. A model of illumination of 3D objects. Let us consider illumination of a three-dimensional object where we know the direction v of the light falling onto the two-dimensional surface of this object, i. e. a set M given implicitly by an equation $F(x,y,z) = 0$. The light intensity of 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 we have seen, the normal line is determined by the gradient of the function F. The sign of our expression then says which side of the surface is illuminated.

For example, consider an illumination with 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 $F(x,y,z) = x^2 + y^2 + z^2 - 1 \le 0$ be the object of our interest. Then, for a point $P = (x,y,z) \in M$ on the surface, we get intensity

$$I(P) = \frac{\operatorname{grad} F\cdot v}{\|\operatorname{grad} F\|\|v\|}I_0 = \frac{-2x - 2y + 2z}{2\sqrt{3}}I_0.$$

We can 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.

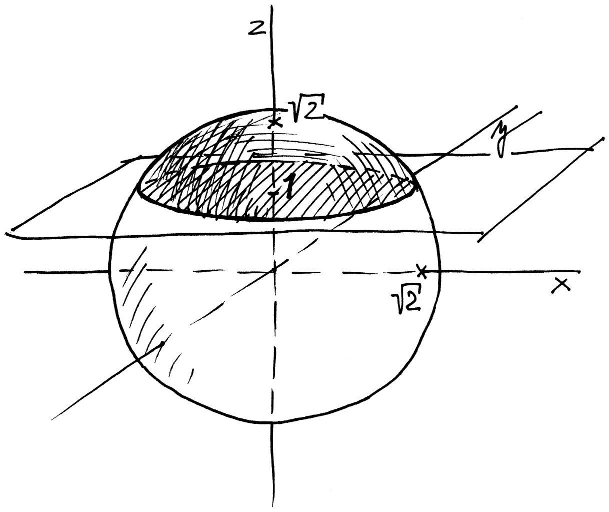
8.21. Tangent and normal spaces. Now, we extend our reasonings about tangent and normal lines to general dimensions. Having a mapping $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$, with coordinate functions f_i, we can also consider the n equations for $n+m$ variables

$$f_i(x_1,\dots,x_{m+n}) = b_i, \quad i = 1,\dots,n,$$

expressing the equality $F(x) = b$ for a vector $b \in \mathbb{R}^n$.

Then, assuming that the conditions of the implicit function theorem hold, the set of all solutions $(x_1,\dots,x_{m+n}) \in \mathbb{R}^{m+n}$ is (at least locally) the graph of a mapping $G : \mathbb{R}^m \to \mathbb{R}^n$.

For a fixed choice $b = (b_1,\dots,b_n)$, the set of all solutions is, of course, the intersection of all hypersurfaces $M(b_i, f_i)$ corresponding to the particular functions f_i. The same must hold for tangent directions, while normal directions are generated by particular gradients. Therefore, if D^1F is the Jacobian matrix of a mapping which implicitly defines a set M and a point $P =$



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

$$V_{\text{sector}} - V_{\text{cone}} = \frac{4}{3}(\sqrt{2}-1) - \frac{1}{3}\pi = \frac{1}{3}\pi(4\sqrt{2}-5).$$

The volume of a general spherical segment with height h in a ball with radius R could be computed similarly:

$$\begin{aligned} V &= V_{\text{sector}} - V_{\text{cone}} \\ &= \int_0^{2\pi}\int_0^{\arccos\left(\frac{R-h}{R}\right)}\int_0^R r^2\sin(\theta)\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &\quad -\frac{1}{3}\pi(2Rh-h^2)(R-h) \\ &= \frac{1}{3}\pi h^2(3R-h). \end{aligned}$$

□

8.72. Find the volume of the part of the cylinder $x^2 + z^2 = 16$ which lies inside the cylinder $x^2 + y^2 = 16$.

$(a_1, \dots, a_{m+n}) \in M$ such that M is the graph of a mapping in the neighborhood of the point P,

$$D^1F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_{m+n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_{m+n}} \end{pmatrix},$$

then the affine subspace in $\mathbb{R}^{m+n}$ which contains exactly all tangent lines going through the point P is given implicitly by the following equations:

$$0 = \frac{\partial f_1}{\partial x_1}(P)\cdot(x_1-a_1) + \dots + \frac{\partial f_1}{\partial x_n}(P)\cdot(x_{m+n}-a_{m+n})$$

$$\vdots$$

$$0 = \frac{\partial f_n}{\partial x_1}(P)\cdot(x_1-a_1) + \dots + \frac{\partial f_n}{\partial x_n}(P)\cdot(x_{m+n}-a_{m+n}).$$

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 $f_1, \dots, f_n$ at the point P, i. e. the rows of the Jacobian matrix D^1F.

As an illustrative simple example, we can calculate the tangent and normal spaces to a conic section in $\mathbb{R}^3$. Let us 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 picture!). Its tangent line at the point P is given by the following equations:

$$\begin{aligned} 0 &= -\frac{1}{2\sqrt{x^2+y^2}}2x\Big|_{x=1,y=0}\cdot(x-1) \\ &\quad -\frac{1}{2\sqrt{x^2+y^2}}2y\Big|_{x=1,y=0}\cdot y + 1\cdot(z-1) \\ &= -x+z \\ 0 &= -2(x-1)+y+(z-1) = -2x+y+z+1, \end{aligned}$$

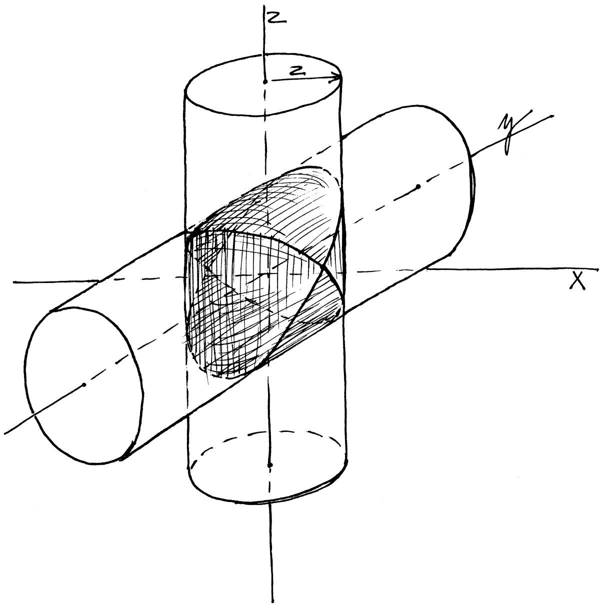
while the plane perpendicular to our curve, containing the point P, is given parametrically by the expression

$$(1, 0, 1) + \tau(-1, 0, 1) + \sigma(-2, 1, 1)$$

with parameters τ and σ.

8.22. Bound extrema. Now, we will approach the first really serious application of the differential calculus of more variables. The typical task of optimization is to find the extrema of values depending on several (yet finitely many) parameters, given some further conditions on the mutual relations between the parameters.

The problem often has $m + n$ parameters which are bound by n conditions. In the language of our differential calculus, we are thus looking for the extrema of a differentiable function h on the set M of points given implicitly by an equation $F(x_1, \dots, x_{m+n}) = 0$. However, we have already prepare efficient procedures for this.

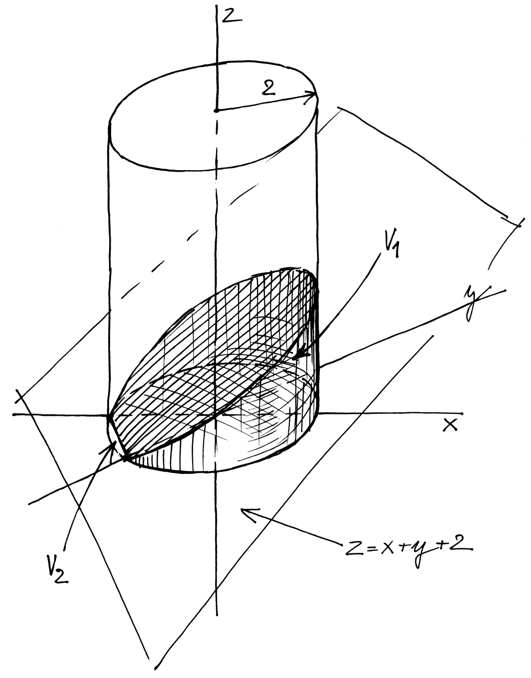


Solution. We will compute the integral in Cartesian coordinates. Since the solid is symmetric, it suffices to integrate over the first octant (interchanging x and $-x$ does not change the equation of the solid; the same holds for y and for z). The part of the solid that lies in the first octant is given by the space under the graph of the function $z(x, y) = \sqrt{16 - x^2}$ and over the quarter-disc $x^2 + y^2 \le 16$, $x \ge 0$, $y \ge 0$. Therefore, the volume of the whole solid is equal to

$$V = 8 \int_0^4 \int_0^{\sqrt{16-x^2}} \frac{4}{\sqrt{16-x^2}} \, \mathrm{d}y \, \mathrm{d}x = 128. \qquad \square$$

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.

8.73. Find the volume of the part of the cylinder $x^2 + y^2 = 4$ bounded by the planes $z = 0$ and $z = x + y + 2$.



For every curve $c(t) \subset M$ going through $P = c(0)$, we must have that $h(c(t))$ is an extremum of this univariate function. Therefore, the derivative must satisfy

$$\frac{d}{dt} h(c(t))_{|t=0} = d_{c'(0)} h(P) = dh(P)(c'(0)) = 0.$$

However, this means that the differential of the function h at the point P is zero along all tangent increases to M at P. This property is equivalent to stating that the gradient of h lies in the normal subspace (more precisely, in its direction space). Such points $P \in M$ are called *stationary points* of the function H with respect to bindings given by F.

As we have seen in the previous paragraph, the normal space to our set M is generated by the rows of the Jacobian matrix of the mapping F, so the stationary points are determined equivalently by the following proposition:

Lagrange multipliers

Theorem. *Let $F = (f_1, \dots, f_n) : \mathbb{R}^{m+n} \to \mathbb{R}^n$ be a differentiable function in a neighborhood 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 conditions F if and only if there exist real parameters $\lambda_1, \dots, \lambda_n$ such that*

$$\operatorname{grad} h = \lambda_1 \operatorname{grad} f_1 + \cdots + \lambda_n \operatorname{grad} f_n.$$

Let us notice that the *method of Lagrange multipliers* is an algorithmic one. Therefore, let us take a look at the numbers of unknowns and equations: the gradients are vectors of $m+n$ coordinates, so the request of the theorem gives $m+n$ equations. The variables are, on one side, the coordinates $x_1, \dots, x_{m+n}$ of the wanted stationary points P with respect to the bindings, and, on the other hand, the n parameters λ_i in the linear combination. Now it remains to state that the point P belongs to the implicitly given set M, which leads to n more equations. Altogether, we have $2n + m$ equations for $2n + m$ variables, so we can expect that the solution will be given by a discrete set of points P (i. e., each one of them will be an isolated point).

8.23. Arithmetic mean–geometric mean inequality. As an example of practical application of the Lagrange multipliers, we will prove the inequality

$$\frac{1}{n}(x_1 + \cdots + x_n) \ge \sqrt[n]{x_1 \cdots x_n}$$

for any n positive real numbers $x_1, \dots, x_n$. Further, we will prove that the inequality holds with equality if and only if all the x_i's are equal.

Let us thus take the sum $x_1 + \cdots + x_n = c$ to be the binding condition for a (non-specified) non-negative constant c. We will look for the maxima and minima of the function

$$f(x_1, \dots, x_n) = \sqrt[n]{x_1 \cdots x_n}$$

with respect to our binding condition and the assumption $x_1 > 0, \dots, x_n > 0$.

The normal vector to the hyperplane defined by the condition is $(1, \dots, 1)$. Therefore, the function f can have an extremum only

Solution. We will work in cylindric coordinates given by the equations $x = r\cos(\varphi)$, $y = r\sin(\varphi)$, $z = z$. The Jacobian of this transformation is $J = r$. The solid can be divided into two parts: above and below the plane $z = 0$, whose volumes will be denoted by V_1 and V_2, respectively. Further, we can notice that one part of the solid with volume V_1 is a pyramid with vertices $[0, 0, 0]$, $[0, 0, 2]$, $[-2, 0, 0]$, $[0, -2, 0]$. Thus, we will further split this solid (above $z = 0$) into two parts, whose volumes we will calculate separately.

$$\begin{aligned} V_1 - V_{\text{pyramid}} &= \int_{-\pi/2}^{\pi} \left(\int_0^2 [r\sin\varphi + r\cos\varphi + 2] r \,\mathrm{d}r \right) \mathrm{d}\varphi, \\ &= 6\pi + \frac{16}{3}, \\ V_{\text{pyramid}} &= \frac{4}{3} \end{aligned}$$

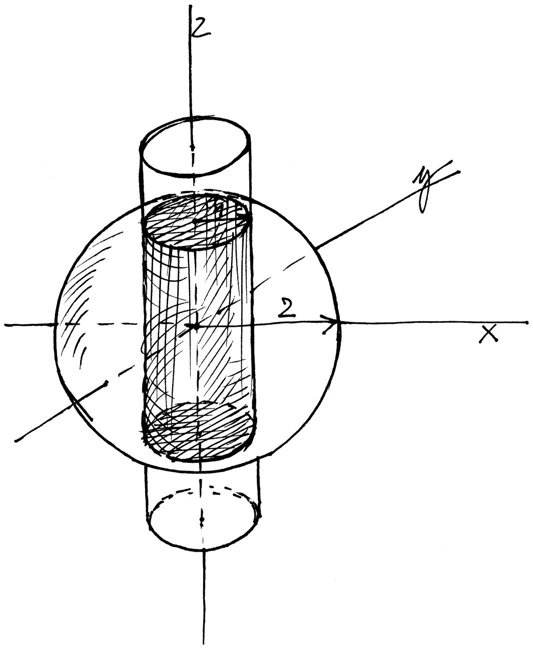
Further,

$$V_1 - V_2 = \int_{-\pi}^{\pi} \int_0^2 r^2(\sin(\varphi) + \cos(\varphi)) + 2r \,\mathrm{d}r \,\mathrm{d}\varphi = 8\pi,$$

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 $\mathbb{R}^2$ 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 $z = 0$.

8.74. 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

at those points where its gradient is a multiple of this normal vector. Thus, we get the following system of equations for the wanted points:

$$\frac{1}{n}\frac{1}{x_i}\sqrt[n]{x_1 \cdots x_n} = \lambda,$$

for $i = 1, \dots, n$ and $\lambda \in \mathbb{R}$.

Apparently, this system has a unique solution $x_1 = \cdots = x_n$ in the examined set. If we allowed the variables x_i to be zero as well, then our set M would be compact, so the function f would have to have both a maximum and a minimum there. However, f is apparently minimal if and only if at least one of the values x_i is zero; so the function necessarily has a strict maximum at our point with $x_i = \frac{c}{n}$, $i = 1, \dots, n$.

The value of the geometric mean is then smaller at all other points with the given sum c of coordinates, which proves the inequality.

2. Integration for the second time

Now, we will return to the process of integration, which was partially described in the second part of chapter six. We will not go into details; instead, we will concentrate on extension of this process quantities dependent on several variables, or on parameters.

8.24. Integrals dependent on parameters. When integrating a function $f(x, y_1, \dots, y_n)$ of $n + 1$ variables with respect to a single variable x, then the result will be a function $F(y_1, \dots, y_n)$ of the remaining variables.

In problems from practice, we often deal with the task to examine such a function F. We can, for instance, look for the volume or area of a body which depends on parameters, and determine the minimal and maximal values (with additional bindings as well). We know from the first part of this chapter that we have tools built upon partial derivatives of functions for this purpose. Therefore, it would be great if we could interchange the operations of differentiation and integration, which we will prove shortly. We begin with examination of continuous dependency upon parameters.

Theorem. *For a continuous function $f(x, y_1, \dots, y_n)$ defined for all x lying in a finite interval $[a, b]$ and all $(y_1, \dots, y_n)$ from a neighborhood U of a point $c = (c_1, \dots, c_n) \in \mathbb{R}^n$, consider the integral*

$$F(y_1, \dots, y_n) = \int_a^b f(x, y_1, \dots, y_n)\, dx.$$

Then the function $F(y_1 \dots, y_n)$ is also continuous in the neighborhood U of the point c.

Proof. To verify the first proposition, it is surely sufficient to recall the definition of the Riemann integral and the already verified fact that a function which is continuous on a compact set is actually uniformly continuous.

Let us thus choose a neighborhood W of a fixed point y so that we would have for all $\bar{y} \in W$ and all $x \in [a, b]$ that

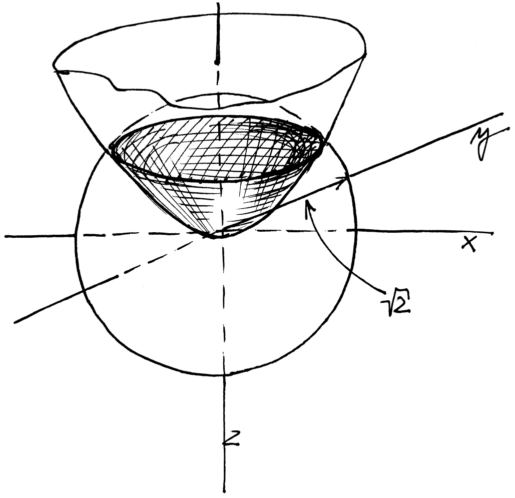
$$|f(x, \bar{y}) - f(x, y)| < \varepsilon$$

for a selected (small) positive ε. The Riemann integral is, for any continuous function, evaluated by approximations by finite sums (equivalently: upper, lower, or Riemann sums with arbitrary representatives ξ_i, see paragraph 6.25 in the sixth chapter). However, can directly write is as the double integral

$$V = 8\int_0^{\pi/2}\int_0^1 r\sqrt{4-r^2}\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{2}{3}(8-3\sqrt{3})\pi.$$

□

8.75. Find the volume of the solid in $\mathbb{R}^3$ which is given by the intersection of the sphere $x^2+y^2+z^2=2$ and the paraboloid $z=x^2+y^2$.

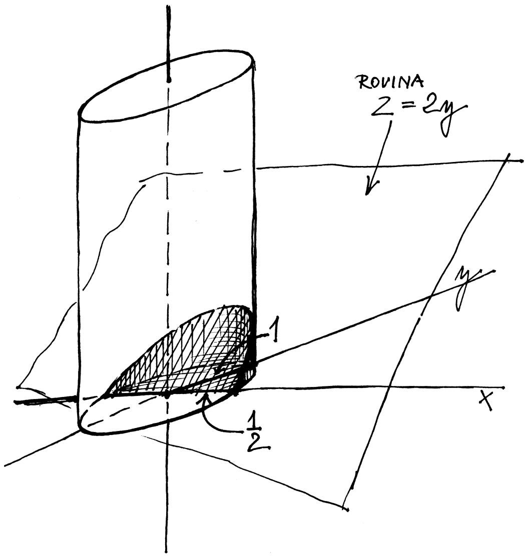


Solution. Once again, we will work in cylindric coordinates:

$$V = \int_0^{2\pi}\int_0^1\int_{r^2}^{\sqrt{2-r^2}} r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{4\sqrt{2}\pi}{3}-\frac{7\pi}{6}.$$

□

8.76. 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 $r^2=1$.

for the chosen parameters $\bar{y}$ and y, none of the sums can differ by more than

$$\begin{aligned}\left|\sum_{i=0}^{k-1} f(\xi_i,\bar{y})(x_{i+1}-x_i)-\sum_{i=0}^{k-1} f(\xi_i,y)(x_{i+1}-x_i)\right|&\\ \le \sum_{i=0}^{k-1}\big|f(\xi_i,\bar{y})-f(\xi_i,y)\big|(x_{i+1}-x_i)&\\ <\varepsilon(b-a).&\end{aligned}$$

Hence it follows that the limit values $F(y)$ and $F(\bar{y})$ cannot differ by more than $\varepsilon(b-a)$ either, so the function is continuous. □

8.25. Integration of multivariate functions. In the case of univariate functions, integration was motivated by the idea of the area under the graph of a given function of one variable. Now, we will consider 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. Back then, we chose small intervals $[x_i,x_{i+1}]$ of length Δx_i which divided the whole interval over which we integrated. Then, we selected their representatives ξ_i, and we approximated the corresponding part of the area by the area of the rectangle with height given by the value $f(\xi_i)$ at the representative, i. e. the expression $f(\xi)\Delta x_i$.

In the case of functions of two variables, we will work with divisions in both and the values representing the height of the graph above the particular little rectangles in the plane.

However, the first thing we must do is to determine the integration domain, i. e. the area we wish to integrate our function f over. As an example, let us take a look at the function $z=f(x,y)=\sqrt{1-x^2-y^2}$, whose graph is, inside the unit disc, the unit sphere. Therefore, 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, we talk about a *multidimensional interval*. If M is a different bounded set in $\mathbb{R}^2$, we work with a sufficiently large area $[a,b]\times[c,d]$, rather than with the set itself, and we adjust our function so that $f(x,y)=0$ for all points lying outside M. Considering the above case of the unit ball, we would thus integrate over the set $M=[-1,1]\times[-1,1]$ the function

$$f(x,y)=\begin{cases}\sqrt{1-x^2-y^2} & \text{for } x^2+y^2\le 1\\ 0 & \text{otherwise.}\end{cases}$$

The definition of the Riemann integral then faithfully follows our procedure from paragraph 6.24. We can do this for an arbitrary finite number of variables.

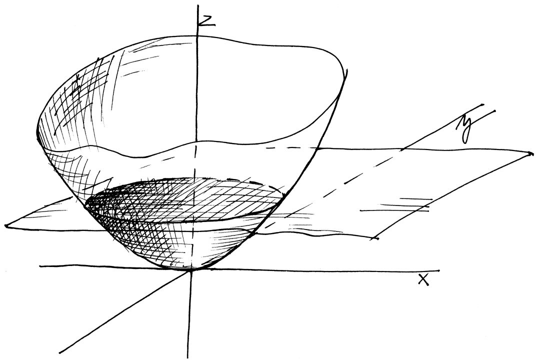
Riemann integral

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 (we are dividing the multidimensional interval in all variables contemporaneously) and the representatives of the particular little cubes $\xi_{i_1\ldots i_n}$, with the maximum size among all used intervals approaching zero, Thus, the wanted volume is

$$\begin{aligned} V &= \int_0^\pi \int_0^1 r\sin(\varphi)\frac{1}{2}r \,\mathrm{d}r\,\mathrm{d}\varphi \\ &= \int_0^\pi \int_0^1 r^2\sin(\varphi)\,\mathrm{d}r\,\mathrm{d}\varphi = \int_0^\pi \frac{1}{3}\sin(\varphi)\,\mathrm{d}\varphi = \frac{2}{3}. \end{aligned}$$

□

8.77. Find the volume of the solid in $\mathbb{R}^3$ which is bounded by the paraboloid $2x^2 + y^2 = z$ and the plane $z = 2$.

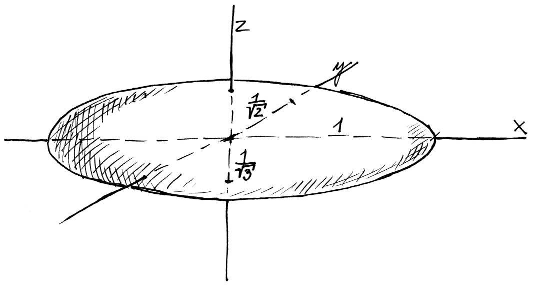


Solution. Similarly to the above problem, we choose "special" coordinates which respect the symmetry of the solid: $x = \frac{1}{\sqrt{2}}r\cos(\varphi)$, $y = r\sin(\varphi)$, $z = z$ with Jacobian $J = \frac{1}{\sqrt{2}}r$. The equation of the paraboloid in these coordinates is $z = r^2$, so the volume of the solid is equal to

$$\begin{aligned} V &= 4\int_0^{\pi/2}\int_0^{\sqrt{2}}\int_{r^2}^{2} \frac{1}{\sqrt{2}}r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi \\ &= 2\sqrt{2}\int_0^{\pi/2}\int_0^{\sqrt{2}} 2r - r^3\,\mathrm{d}r\,\mathrm{d}\varphi = 2\sqrt{2}\int_0^{\pi/2}\mathrm{d}\varphi \\ &= \sqrt{2}\pi. \end{aligned}$$

□

8.78. Calculate the volume of the ellipsoid $x^2 + 2y^2 + 3z^2 = 1$.



Solution. We will consider the coordinates

$$\begin{aligned} x &= r\cos(\varphi)\sin(\theta), \\ y &= \frac{1}{\sqrt{2}}r\sin(\varphi)\sin(\theta), \\ z &= \frac{1}{\sqrt{3}}r\cos(\theta). \end{aligned}$$

The corresponding Jacobian is $\frac{1}{\sqrt{6}}r^2\sin(\theta)$, so the volume is

$$V = \int_0^{2\pi}\int_0^{\pi}\int_0^1 \frac{1}{\sqrt{6}}r^2\sin(\theta)\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi = \frac{4}{3\sqrt{6}}\pi.$$

the integral sums

$$S_{\Xi,\xi} = \sum_{i_1\dots i_n} f(\xi_{i_1\dots i_n})\Delta x_{i_1\dots i_n}$$

(here, we write $\Delta x_{i_1\dots i_n}$ for the product of the sizes of the particular intervals from the division defining the little cube with the corresponding indeces) always converge to the value

$$S = \int_S f(x_1,\dots,x_n)\,dx_1\dots dx_n,$$

independent of the selected sequence of divisions and representatives.

The function f is then said to be Riemann-integrable over I.

As a relatively simple exercise, you can 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: we control the norms of the divisions used in the definition somewhat roughly.

The situation is much worse if we try to integrate in this way over unbounded intervals, because, unlike integrals in one variable, we cannot replace the wanted result uniquely with the limit of integrals over bounded areas, see **??** below. Therefore, we will further talk about integration of functions over $\mathbb{R}^n$ only for functions whose support is compact, i. e. functions which take zero outside a bounded interval I.

A bounded set $M \subset \mathbb{R}^n$ is said to be *Riemann measurable* iff its indicator function, defined by

$$\chi_M(x_1,\dots,x_n) = \begin{cases} 1 & \text{for } (x_1,\dots,x_n) \in S \\ 0 & \text{for all other points in } \mathbb{R}^n, \end{cases}$$

is Riemann-integrable over $\mathbb{R}^n$.

For any Riemann-measurable set M and a function f defined at all points of M, we can consider the function $\tilde{f} = \chi_M \cdot f$ as a function defined on the whole $\mathbb{R}^n$, and this function $\tilde{f}$ apparently has a compact support. The Riemann integral of the function f over the set M is defined by

$$\int_M f\,dx_1\dots dx_n = \int_{\mathbb{R}^n} \tilde{f}\,dx_1\dots dx_n,$$

supposing the right-hand integral exists.

This definition of the Riemann integral does not provide reasonable instructions how to compute the values of integrals. However, it immediately leads to the following basic properties of the Riemann integral (cf. Theorem 6.24):

8.26. Theorem. *The set of Riemann-integrable real-valued functions over an interval $I \subset \mathbb{R}^n$ is a vector space over the real scalars, and the Riemann integral is a linear form there.*

If the integration domain S is given as a disjoint union of finitely many Riemann-measurable domains S_i, the integral over a function f over S is given by the sums of the integrals over the particular domains S_i.

Proof. All the properties follows directly from the definition of the Riemann integral and the properties of convergent sequences of real numbers, just like in the case of univariate functions. We advise to think out the details by yourselves. □

Now, let us rewrite the theorem into usual equalities:

□

8.79. Remark. Note that if the transformation the coordinates is linear (and affine), then the space is deformed "uniformly". This means that the volume of an arbitrary solid is changed proportionally to the change of the volume of an infinitesimal volume element, which is the Jacobian. Therefore, if we consider the volume of the ball with a given radius r to be known, (in this case, $r = 1$), we can infer directly that the volume of the ellipsoid is $V = \frac{1}{\sqrt{6}} \cdot \frac{4}{3}\pi = \frac{4}{3\sqrt{6}}\pi$.

8.80. 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

$$V = \int_0^{2\pi}\int_0^1\int_{r^2}^1 \frac{r}{\sqrt{10}}\, \mathrm{d}z\, \mathrm{d}r\, \mathrm{d}\varphi = \frac{\pi}{2\sqrt{10}}.$$

□

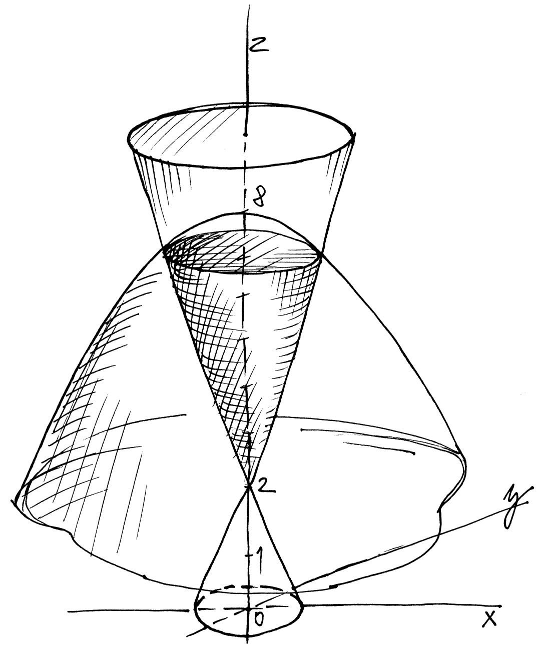
8.81. Find the volume of the solid which lies in the first octant and is bounded by the surfaces $y^2 + z^2 = 9$ and $y^2 = 3x$.

Solution. In cylindric coordinates,

$$V = \int_0^{\pi/2}\int_0^3\int_0^{\frac{r^2}{3}\cos^2(\varphi)} r\, \mathrm{d}x\, \mathrm{d}r\, \mathrm{d}\varphi = \frac{27}{16}\pi.$$

□

8.82. Find the volume of the solid in $\mathbb{R}^3$ which is bounded by the cone part $2x^2 + y^2 = (z-2)^2$, $z \geq 2$ and the paraboloid $2x^2 + y^2 = 8 - z$.



Solution. First of all, we find the intersection of the given surfaces:

$$(z-2)^2 = -z + 8, \ z \geq 2;$$

Finite additivity and linearity

The first part says that a linear combination (over scalars in $\mathbb{R}$) of Riemann-integrable functions $f_i : \mathbb{I} \to \mathbb{R}$, $i = 1, \dots, k$ is always a Riemann-integrable function, and it can be computed as follows:

$$\begin{aligned} &\int_I \Big(a_1 f_1(x_1, \dots, x_n) + \dots + a_k f_k(x_1, \dots, x_n) \Big)\, dx_1 \dots dx_n \\ &\quad = a_1 \int_I f_1(x_1, \dots, x_n)\, dx_1 \dots dx_n + \\ &\qquad \dots + a_k \int_I f_k(x_1, \dots, x_n)\, dx_1 \dots dx_n. \end{aligned}$$

The second part then says that for disjoint Riemann-measurable sets M_1 and M_2 and for a function $f : \mathbb{R}^n \to \mathbb{R}$ which is Riemann-integrable over both these sets, we have that

$$\begin{aligned} &\int_{M_1 \cup M_2} f(x_1, \dots, x_n)\, dx_1 \dots dx_n \\ &\quad = \int_{M_1} f(x_1, \dots, x_n)\, dx_1 \dots dx_n + \\ &\qquad \int_{M_2} f(x_1, \dots, x_n)\, dx_1 \dots dx_n. \end{aligned}$$

8.27. Multiple integrals. We will see in a while that Riemann-integrable functions especially involve the cases when the integration domain M can be defined by a continuous function dependency of the coordinates of boundary points so that, given the first coordinate x, we can define the range of the next coordinate by two functions, i. e., $y \in [\varphi(x), \psi(x)]$, then the range of the next coordinate by $z \in [\eta(x, y), \zeta(x, y)]$, and so on for all of the other coordinates.

We can indeed do this in the case of our ball from the introductory example: for $x \in [-1, 1]$, we 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 we can integrate the indicator function of the ball, i. e. the function which takes one on the area $S \subset \mathbb{R}^3$ which is defined by $z \in [-\sqrt{1-x^2-y^2}, \sqrt{1-x^2-y^2}]$.

The following theorem is fundamental for this. It transforms the computation of a Riemann integral to a gradual computation of several univariate integrals (while the other variables are considered to be parameters, which can thus appear in the integration bounds as well).

Multiple integrals

Theorem. *Let $M \subset \mathbb{R}^n$ be a bounded set given, as above, by continuous functions ψ_i, η_i*

$$\begin{aligned} M = \{(x_1, \dots, x_n);\ & x_1 \in [a, b], x_2 \in [\psi_2(x_1), \eta_2(x_1)], \dots, \\ & x_n \in [\psi_n(x_1, \dots, x_{n-1}), \eta_n(x_1, \dots, x_{n-1})]\}, \end{aligned}$$

and 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 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

$$V = \int_0^{2\pi} \int_0^2 \int_{r+2}^{8-r^2} \frac{r}{\sqrt{2}} \, \mathrm{d}z \, \mathrm{d}r \, \mathrm{d}\varphi = \frac{16\sqrt{2}}{3}\pi.$$

□

8.83. 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,

$$V = \int_0^{2\pi} \int_0^2 \int_0^{8-\frac{r^2}{2}} r \, \mathrm{d}x \, \mathrm{d}r \, \mathrm{d}\varphi = 28\pi.$$

□

8.84. 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:

$$x_t = \iiint_T x \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z,$$

$$y_t = \iiint_T y \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z,$$

$$z_t = \iiint_T z \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z.$$

The centroid of a figure in $\mathbb{R}^2$ or other dimensions can be computed analogously.

8.85. Find the centroid of the part of the ellipse $3x^2 + 2y^2 = 1$ which lies in the first quadrant of the plane $\mathbb{R}^2$.

Solution. First, let us calculate the volume of the given ellipse. The transformation $x = \frac{1}{\sqrt{3}}x'$, $y = \frac{1}{\sqrt{2}}y'$ with Jacobian $\frac{1}{\sqrt{6}}$ leads to

$$S = \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} \mathrm{d}y \, \mathrm{d}x = \frac{1}{\sqrt{6}} \int_0^1 \int_0^{\sqrt{1-x^2}} \mathrm{d}y' \, \mathrm{d}x' = \frac{\pi}{4\sqrt{6}}.$$

The other integrals we need can be computed directly in Cartesian coordinates x and y:

$$\begin{aligned} T_x &= \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} x \, \mathrm{d}y \, \mathrm{d}x = \int_0^{\frac{1}{\sqrt{3}}} x\sqrt{\frac{1-3x^2}{2}} \, \mathrm{d}x = \\ &= \frac{1}{2}\int_0^{\frac{1}{3}} \sqrt{\frac{1-3t}{2}} \, \mathrm{d}t = \frac{\sqrt{2}}{18}, \end{aligned}$$

formula

$$\int_M f(x_1, x_2, \dots, x_n)\, dx_1 \dots dx_n = \int_a^b \Bigg(\int_{\psi_2(x_1)}^{\eta_2(x_1)} \dots$$

$$\Bigg(\int_{\psi_n(x_1,\dots,x_{n-1})}^{\eta_n(x_1,\dots,x_{n-1})} f(x_1, x_2, \dots, x_n)\, dx_n \Bigg) \dots dx_2 \Bigg) dx_1$$

PROOF. First of all, we will go through the proof for the case of two variables, and then we will see that there is no need of further ideas in the general case.

Consider an interval $I = [a, b] \times [c, d]$ containing our 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

$$\begin{aligned} S_{\Xi,\xi} &= \sum_{i,j} f(\xi_{ij})\Delta x_{ij} \\ &= \sum_i \Big(\sum_j f(\xi(ij))\Delta y_j \Big) \Delta x_i, \end{aligned}$$

where we write Δx_{ij} for the product of the sizes Δx_i and Δx_j of the intervals which correspond to the choice of the representative ξ_{ij}.

Now, let us assume that we work only with choices of representatives ξ_{ij} which all share the same first coordinate x_i. If we leave the division of the interval $[a, b]$ and refine only the division of $[c, d]$, the values of the inner sum of our expression will approach the value of the integral

$$S_i = \int_{\varphi(x_i)}^{\eta(x_i)} f(x_i, y)\, dy,$$

which surely exists since the function $f(x_i, y)$ is continuous. Moreover, we thus obtain a function which is continuous in the free parameter x_i, see 8.24. Therefore, further refinement of the division of the interval $[a, b]$ leads, in limit, to the desired formula

$$\sum_i S_i \Delta x_i \to S = \int_a^b \Bigg(\int_{\psi(x)}^{\eta(y)} f(x, y)\, dy \Bigg) dx.$$

It remains to deal with the case of general choices of representatives of general divisions Ξ. However, since we are working with a continuous function f on a compact set, it is actually uniformly continuous there. Therefore, if we select a small real number $\varepsilon > 0$ beforehand, we can always, for the norm of a division, find a bound $\delta > 0$ so that the values of the function f for the general choices x_{ij} differs by no more than ε from the choices used above. The limit processes thus result in the same for general Riemann sums $S_{\Xi,\xi}$ as we saw above.

Now, the general case can be proved easily by induction. In the case of $n = 1$, the result is trivial. The presented reasoning can easily be transformed for a general induction step, writing $(x_2, \dots, x_n)$ instead of y, having x_1 instead of x, and perceiving the particular little cubes of the divisions as $(n-1)$-dimensional cubes Cartesian-multiplied by the last interval. In the last-but-one step of the proof, we just use the induction hypothesis, rather than the simple one-dimensional integration. The final argument about uniform continuity remains

$$\begin{aligned} T_y &= \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} y \,\mathrm{d}y \,\mathrm{d}x = \frac{1}{2}\int_0^{\frac{1}{\sqrt{3}}} \frac{1-3x^2}{2}\,\mathrm{d}x = \\ &= \frac{1}{4}\int_0^{\frac{1}{\sqrt{3}}} (1-3x^2)\,\mathrm{d}x = \frac{\sqrt{3}}{18}. \end{aligned}$$

Therefore, the coordinates of the centroid are $[\frac{4\sqrt{3}}{9\pi}, \frac{2\sqrt{2}}{\pi}]$. □

8.86. 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

$$V = 4\int_0^{\pi/2}\int_0^r \int_{\frac{h}{r}\rho}^h \rho \,\mathrm{d}z\,\mathrm{d}\rho\,\mathrm{d}\varphi = \frac{1}{3}\pi h r^2.$$

Apparently, the centroid lies on the z-axis. For the z-coordinate, we get

$$z = \frac{1}{V}\int_{\text{kuĹžel}} z dV = \frac{1}{V}\int_0^{\pi/2}\int_0^r\int_{\frac{h}{r}\rho}^h z\rho\,\mathrm{d}z\,\mathrm{d}\rho\,\mathrm{d}\varphi = \frac{3}{4}h.$$

Thus, the centroid lies $\frac{1}{4}h$ over the center of the cone's base. □

8.87. Find the centroid of the solid which is bounded by the paraboloid $2x^2 + 2y^2 = z$, the cylinder $(x+1)^2 + y^2 = 0$, and the plane $z = 0$.

Solution. First, we will compute the volume of the given solid. Again, we use the cylindric coordinates ($x = r\cdot\cos\varphi$, $y = r\cdot\sin\varphi$, $z = z$), where the equation of the paraboloid is $z = 2r^2$ and the equation of the cylinder reads $r = -2\cos(\varphi)$. Moreover, taking into account the fact that the plane $x = 0$ is tangent to the given cylinder, we can easily determine the bounds of the integral that corresponds to the volume of the examined solid:

$$\begin{aligned} V &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_0^{-2\cos\varphi}\int_0^{2r^2} r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi \\ &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_0^{-2\cos\varphi} 2r^3\,\mathrm{d}r\,\mathrm{d}\varphi \\ &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} 8\cos^4\pi\varphi = 3\pi, \end{aligned}$$

where the last integral can be computed using the method of recurrence from 6.22.

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

the same. We advise to go through this proof in detail as an exercise. □

Fubini's theorem

8.28. Corollary. *For a multidimensional interval $M = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ and a continuous function $f(x_1, \ldots, x_n)$ on M, the multiple Riemann integral*

$$\begin{aligned} &\int_M f(x_1,\ldots,x_n)\,dx_1\ldots dx_n \\ &\quad = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\ldots\int_{a_n}^{b_n} f(x_1,\ldots,x_n)\,dx_1\ldots dx_n \end{aligned}$$

is independent of the order in which the integrations are performed.

Proof. In the case of a multidimensional interval M in the previous theorem, any order of integration expresses the area M in the required form. Therefore, the order of integration has no effect upon the result of the integral. □

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 connection of Fourier transforms and convolutions, see paragraph 7.9.

Our simple derivation of Fubini's 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 we have just introduced was used by mathematicians like Cauchy at least a century before Fubini.

We can also notice that we have defined no concept of an improper integral for unbounded multivariate functions. You can verify that it is quite impossible to do this in a reasonable way, just consider the following example of two multiple integrals:

$$\begin{aligned} \int_0^1\Big(\int_0^1 \frac{x-y}{(x+y)^3}\,dy\Big)dx &= \frac{1}{2} \\ \int_0^1\Big(\int_0^1 \frac{x-y}{(x+y)^3}\,dx\Big)dy &= -\frac{1}{2}. \end{aligned}$$

The reason can be felt already from the properties of non-absolutely converging series. There, rearranging the summands can lead to an arbitrary result.

The situation is a bit better if we calculate the Riemann integral of a bounded Riemann-integrable function $f(x)$ with non-compact support over the whole $\mathbb{R}^n$. If there is a universal bound

$$\left|\int_I f(x)\,dx\right| \le C$$

with a constant C independent of the choice of an n–dimensional interval I, then it is possible to define

$$\int_{\mathbb{R}^n} f(x)\,dx = \lim_{r\to\infty}\int_{I_r} f(x)\,dx,$$

where $I_r = \{(x_1,\ldots,x_n);\ |x_j| < r,\ j = 1,\ldots,n\}$, and the result is, of course, bounded by the same constant C. In this case as well, be computed by the following integrals:

$$\begin{aligned} x_T &= \frac{1}{V}\int\int\underset{\text{B}}{\int} x\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \frac{1}{V}\int_0^{2r^2}\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_0^{-2\cos\varphi} r^2\cos\varphi\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\varphi \\ &= \frac{1}{V}\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_0^{-2\cos\varphi} 2r^4\cos\varphi\,\mathrm{d}r\,\mathrm{d}ph \\ &= \frac{1}{V}\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} -\frac{64}{5}\cos^6\varphi\,\mathrm{d}\varphi = -\frac{4}{3}, \end{aligned}$$

where the last integral was computed by 6.22 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 $[-\frac{4}{3}, 0, \frac{20}{9}]$. □

8.88. 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 in exercise ||8.86|| 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 $[0, 0, \frac{3}{2}]$. □

8.89. 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:

$$V = \int_0^{2\pi}\int_0^1\int_{r+2}^{4-r^2} r\,\mathrm{d}z\ \mathrm{d}r\,\mathrm{d}\varphi = \frac{5}{6}\pi.$$

□

8.90. Find the volume of the solid in $\mathbb{R}^3$ which lies under the cone $x^2 + y^2 = (z-2)^2$, $z \le 2$ and over the paraboloid $x^2 + y^2 = z$.

Solution.

$$V = \int_0^{2\pi}\int_0^1\int_{r^2}^{2-r} r\,\mathrm{d}z\ \mathrm{d}r\,\mathrm{d}\varphi = \frac{5}{6}\pi.$$

Note that the considered solid is symmetric with the solid from the previous exercise ||8.89|| (the center of the symmetry is the point [0, 0, 2]). Therefore, it must have the same volume. □

Fubini's theorem holds in the form

$$\int_{\mathbb{R}^n} f(x)\,dx = \int_{-\infty}^{\infty}\dots\left(\int_{-\infty}^{\infty} f(x)\,dx_1\right)\dots dx_n.$$

8.29. Notes about integration. The Riemann integral of multivariate functions behaves even worse than we have seen in the case of functions of one variable in the sixth chapter. Therefore, more sophisticated approaches to integrations have been developed, which are derived from the concept of the measure of a set. Let us take a quick look at this problem.

We can consider the strict analogy of the lower and upper Riemann integrals for univariate functions. This means taking infima or suprema, respectively, over the corresponding multidimensional interval, instead of the function values at the representatives in the Riemann sums. For bounded functions, we always get well-defined values this way, and if we do this for the indicator function χ_M of a fixed set M, we get the so-called inner and outer Riemann measure of the set M. Apparently, the inner measure is the limit of the areas given by the sum of the volumes of all intervals from our divisions which are inside M, and, on the other hand, the outer measure is given by the sum 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 lower and upper measures are equal.

The sets whose outer measure is zero are, of course, Riemann-measurable. We call them measure-zero sets or null sets. It can be shown quite easily that Riemann-integrable functions are exactly those bounded functions with compact support whose set of discontinuity points has measure zero. Surely, this definition makes the measure finitely additive, i. e., 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 sets being united. However, unlike in the case of one variable, now it does not hold that a countable disjoint union of measurable sets is measurable, so we must expect problems with limit approaches, as we have seen in the case of one variable.

If we restrict ourselves to the vector space $\mathcal{S}_c(\mathbb{R}^n)$ of all continuous functions with compact support, we can proceed in the same way as in the seventh chapter, i. e., we can define, for functions $f \in \mathcal{S}_c(\mathbb{R}^n)$, their norms

$$\|f\|_p = \left(\int_{\mathbb{R}^n} |f(x_1,\dots,x_n)|^p\,dx_1\dots dx_n\right)^{1/p}$$

for all values $1 \le p < \infty$. Thanks to the Riemann integral having been defined in terms of divisions, the properties of the norm can be verified in the same way as for univariate functions, using Hölder's and Minkowski's inequalities.

We thus get the metric spaces $\mathcal{L}^p$. As we have known from the general theory, its completion exists (and it is determined uniquely, up to isometry), and it can be shown that it will again be a space of functions. Moreover, a more general theory of integration can be developed so that the norms on these complete spaces would be given by the same formulae as above. However, we will not go deeper into these parts of mathematical analysis here.

8.91. Find the centroid of the surface bounded by the parabola $y = 4 - x^2$ and the line $y = 0$. ○

8.92. Find the centroid of the circular sector corresponding to the angle of 60° that was cut out of a disc with radius 1. ○

8.93. Find the centroid of the semidisc $x^2 + y^2 = 1$, $y \geq 0$. ○

8.94. Find the centroid of the circular sector corresponding to the angle of 120° that was cut out of a disc with radius 1. ○

8.95. Find the volume of the solid in $\mathbb{R}^3$ which is given by the inequalities $z \geq 0$, $z - x \leq 0$, and $(x-1)^2 + y^2 \leq 1$. ○

8.96. Find the volume of the solid in $\mathbb{R}^3$ which is given by the inequalities $z \geq 0$, $z - y \leq 0$. ○

8.97. Find the volume of the solid bounded by the surface

$$3x^2 + 2y^2 + 3z^2 + 2xy - 2yz - 4xz = 1.$$

○

8.98. 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 $x \geq 1$. ○

8.99. 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

$$\begin{aligned} &\iint_K \sqrt{1 + f_x^2 + f_y^2}\,\mathrm{d}x\,\mathrm{d}y = \iint_K \sqrt{2}\,\mathrm{d}x\,\mathrm{d}y = \\ = \ &\sqrt{2}\int_0^{\pi}\int_0^{\sin\pi} r\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{\sqrt{2}}{2}\int_0^{\pi} \sin^2\varphi \\ = \ &\frac{\sqrt{2}\pi}{4}. \end{aligned}$$

□

8.100. Find the area of the parabola $z = x^2 + y^2$ over the disc $x^2 + y^2 \leq 4$. ○

8.101. 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.

8.30. Differentiation with respect to parameters. Now, we can finally deal with the promised dependency of integrals upon parameters. The following result is highly applicable. For instance, we can use it when examining integral transforms, which we talked about in the second part of chapter seven.

Now, our 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.

Differentiation with respect to parameters

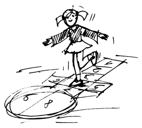
Theorem. *For a continuous function $f(x, y_1, \dots, y_n)$ defined for all x from a finite interval $[a, b]$ and for all $(y_1, \dots, y_n)$ lying in some neighborhood U of a point $c = (c_1, \dots, c_n) \in \mathbb{R}^n$, consider the integral*

$$F(y_1, \dots, y_n) = \int_a^b f(x, y_1, \dots, y_n)\,dx.$$

If there exists a continuous partial derivative $\frac{\partial f}{\partial y_j}$ on a neighborhood of the point c, then $\frac{\partial F}{\partial y_j}(c)$ exists as well, and we have

$$\frac{\partial F}{\partial y_j}(c) = \int_a^b \frac{\partial f}{\partial y_j}(x, c_1, \dots, c_n)\,dx.$$

Proof. Thanks to the considered continuity of all functions, we can easily utilize our knowledge about univariate antiderivatives, and the result will be a simple consequence of Fubini's theorem. Since all the other parameters y_j play only the passive role of a constant parameter in our reasonings, we can assume without loss of generality that there is only one parameter y.



Let us denote

$$G(y) = \int_a^b \frac{\partial f}{\partial y}(x, y)\,dx, \quad F(y) = \int_a^b f(x, y)\,dx$$

and compute, invoking Fubini's theorem, the antiderivative

$$\begin{aligned} H(y) &= \int_{y_0}^{y} G(y)\,dy = \int_a^b \Big(\int_{y_0}^{y} \frac{\partial f}{\partial y}(x, y)\,dy\Big)\,dx \\ &= F(y) - F(y_0). \end{aligned}$$

Finally, differentiating with respect to y yields

$$G(y) = \frac{\partial H}{\partial y}(y) = \frac{\partial F}{\partial y}(y),$$

which is what we have wanted to prove. □

8.31. Change of coordinates at integration. When calculating integrals of univariate functions, we used coordinate transformations as an extraordinarily powerful tool. The situation is very similar in the case of functions of more variables.

First, let us recall (with an appropriate interpretation for the subsequent generalization) the transformation for a single variable. The integrated expression $f(x)\,dx$ describes the area of a rectangle defined by a (linearized) increase of the variable x and by the value

8.102. 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 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):

$f(x)$. If we transform the variable by a relation $x = u(t)$, then the linearized increase can be expressed as

$$dx = \frac{du}{dt}dt,$$

and so the corresponding contribution for the integral is given by

$$f(u(t))\frac{du}{dt}dt,$$

where we either suppose that the sign of the derivative $u'(t)$ is positive, or we interchange the bounds of the integral, so that the sign takes no effect in the result.

Intuitively, the procedure for n variables is quite similar. We only have to use our knowledge about the volume of parallelepipeds from linear algebra.

We use, for Riemann integrals in the Riemann sums, an approximation which takes the volume (area) of a small multidimensional interval and multiplies it by the value of the function at the representative point. If we transform the coordinates, we not only get the function value at the representative point in a new coordinate expression, but we also have to account for the change of the area or volume of the corresponding small multidimensional interval. Once again, this is the case of a linear approximation of a change, which we know well — this is actually an action of the linear approximation of the used transformation, i. e. an action of the Jacobian matrix, see 8.14. The change of the volume is then given (in absolute value) by the determinant of this matrix (see our discussion of this topic in linear algebra, especially 4.22).

Transformation of coordinates

Theorem. *Let $G(t_1,\dots,t_n) : \mathbb{R}^n \to \mathbb{R}^n$, $(x_1,\dots,x_n) = G(t_1,\dots,t_n)$, be a continuously differentiable mapping, $N = G(M)$ and M be Riemann-measurable sets, and $f : M \to \mathbb{R}$ a continuous function. Then,*

$$\begin{aligned}
&\int_M f(x_1,\dots,x_n)\,dx_1\dots dx_n = \\
&\quad = \int_N f(G(t_1,\dots,t_n))\big|\det(D^1G(t_1,\dots,t_n))\big|\,dt_1\dots dt_n.
\end{aligned}$$

Proof. Since we are working with a continuous function f and a differentiable change of coordinates, the integrals on both sides of the equality to be proved apparently exist. Therefore, we only need to prove that their values are indeed equal.

Let us denote our composite function by

$$g(t_1,\dots,t_n) = f(G(t_1,\dots,t_n)),$$

and choose a sufficiently large n-dimensional interval I containing N and its division Ξ. The entire proof is nothing more than a more exact writing of the discussion presented above.

First, let us notice two things: The images of the boundaries of our interval $I_{i_1\dots i_n}$ are differentiable objects (sides, edges, etc.); in particular, they will again be Riemann-measurable sets. For each little part $I_{i_1\dots i_n}$ of our division Ξ, the integral f over $J_{i_1\dots i_n} = G(I_{i_1\dots i_n})$ surely exists.

Further, if we fix the center $t_{i_1\dots i_n}$ of the interval $I_{i_1\dots i_n}$, then we get the linear image of this interval (note that we map the interval

$$\psi(\xi,\eta)=A\iint_{\mathbb{R}^2} s(x,y)e^{-ik(\xi x+\eta y)}\,\mathrm{d}x\,\mathrm{d}y$$

.

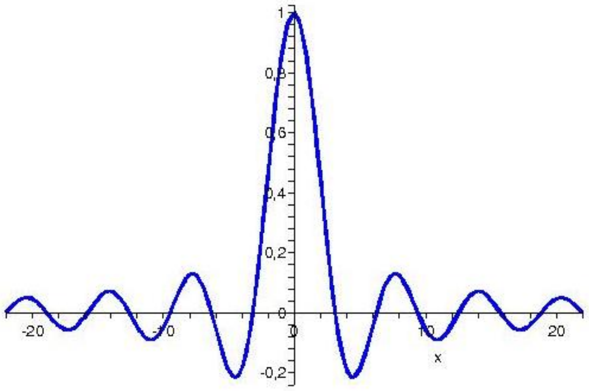
$$\psi(\xi,\eta)=A\int_{-p/2}^{p/2}\int_{-q/2}^{q/2} e^{-ik(\xi x+\eta y)}\,\mathrm{d}y\,\mathrm{d}x$$

$$\psi(\xi,\eta)=A\int_{-p/2}^{p/2} e^{-ik\xi x}\,\mathrm{d}x\int_{-q/2}^{q/2} e^{-ik\eta y}\,\mathrm{d}y=$$

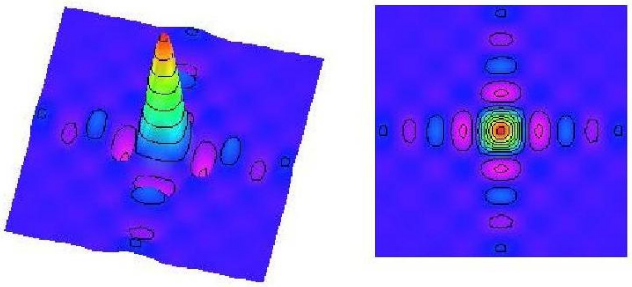
$$=A\left[\frac{e^{-ik\xi x}}{-ik\xi}\right]_{-p/2}^{p/2}\left[\frac{e^{-ik\eta y}}{-ik\eta}\right]_{-q/2}^{q/2}=$$

$$=A\,\frac{2\sin(k\;\xi p/2)}{k\xi}\,\frac{2\sin(k\;\eta q/2)}{k\eta}=Apq\frac{\sin(k\;\xi p/2)}{k\xi p/2}\,\frac{\sin(k\;\eta q/2)}{k\eta q/2}$$

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:

shifted to the origin with the linear mapping given by the Jacobian matrix, and the result is then added to the image of the center)

$$R_{i_1\dots i_n}=G(t_{i_1\dots i_n})+D^1G(t_{i_1\dots i_n})(I_{i_1\dots i_n}-t_{i_1\dots i_n}),$$

an n-dimensional parallelepiped. If our division is very fine, this parallelepiped differs only a bit from the image $J_{i_1,\dots,i_n}$. Exactly speaking, thanks to the uniform continuity of the mapping G, we can, for an arbitrarily small $\varepsilon>0$, find a norm of the division such that we will have for all finer divisions that

$$G(t_{i_1\dots i_n})+(1+\varepsilon)D^1G(t_1,\dots t_n)(I_{i_1\dots i_n})\supset J_{i_1\dots i_k}.$$

However, then the n-dimensional volumes will also satisfy

$$\begin{aligned}\operatorname{vol}_n(J_{i_1\dots i_n})&\le(1+\varepsilon)^n\operatorname{vol}_n(R_{i_1\dots i_n})\\&=(1+\varepsilon)^n\left|\det G(t_{i_1\dots i_k})\right|\operatorname{vol}_n(I_{i_1\dots i_n}).\end{aligned}$$

Now, we are able to bound the whole integral from above:

$$\begin{aligned}&\int_M f(x_1,\dots,x_n)\,dx_1\dots dx_n=\\&\quad=\sum_{i_1\dots i_n}\int_{J_{i_1\dots in}} f(x_1,\dots,x_n)\,dx_1\dots dx_n\\&\quad\le\sum_{i_1\dots i_n}(\sup_{(t_1,\dots,t_n)\in I_{i_1\dots in}} g)\operatorname{vol}_n(J_{i_1\dots i_n})\\&\quad\le(1+\varepsilon)^n\sum_{i_1\dots i_n}(\sup_{(t_1,\dots,t_n)\in I_{i_1\dots in}} g)\left|\det G(t_{i_1\dots i_k})\right|\operatorname{vol}_n(I_{i_1\dots i_n}).\end{aligned}$$

Letting the norms of the divisions approach zero, the left-hand value remains the same, while on the right side, we obtain the Riemann integral. Instead of the equality to be proved, we get the inequality:

$$\begin{aligned}&\int_M f(x_1,\dots,x_n)\,dx_1\dots dx_n\\&\quad\le\int_N f(G(t_1,\dots,t_n))\left|\det(D^1G(t_1,\dots,t_n))\right|dt_1\dots dt_n.\end{aligned}$$

However, now we can repeat the same reasoning so that we interchange G and G^{-1}, the integration domains M and N, and the functions f and g. We thus immediately obtain the other inequality:

$$\begin{aligned}&\int_N g(t_1,\dots,t_n)\left|\det(D^1G(t_1,\dots,t_n))\right|dt_1\dots dt_n\\&\quad\le\int_M f(x_1,\dots,x_n)\left|\det(D^1G(G^{-1}(x_1,\dots,x_n)))\right|\\&\qquad\quad\left|\det(D^1G^{-1}(x_1,\dots,x_n))\right|dx_1\dots dx_n\\&\quad=\int_M f(x_1,\dots,x_n))\,dx_1\dots dx_n,\end{aligned}$$

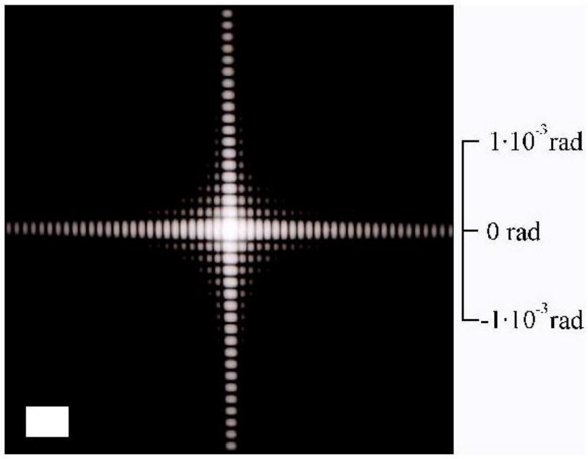
which finishes the proof. □

8.32. 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 $g(s,t)=f(x(s,t),y(s,t))$, we get

$$\int_{G(N)} f(x,y)\,dxdy=\int_N g(s,t)\left|\frac{\partial x}{\partial s}\frac{\partial y}{\partial t}-\frac{\partial x}{\partial t}\frac{\partial y}{\partial s}\right|dsdt.$$



Since $\lim_{x\to 0}\frac{\sin x}{x} = 1$, the intensity at the middle of the image is directly proportional to $I_0 = A^2p^2q^2$. 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.

I. Applications of Stoke's theorem – Green's theorem

8.103. 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 8.44), 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:

$$\begin{aligned}\int_c (x-y)\ \mathrm{d}x + x\ \mathrm{d}y &= \iint_D (1+1)\ \mathrm{d}x\ \mathrm{d}y == 2\iint_D \mathrm{d}x\ \mathrm{d}y = \\ &= 2\int_{-2}^{2}\int_{-2}^{2} \mathrm{d}x\ \mathrm{d}y = 2\big[x\big]_{-2}^{2}\cdot\big[y\big]_{-2}^{2} = 32.\end{aligned}$$

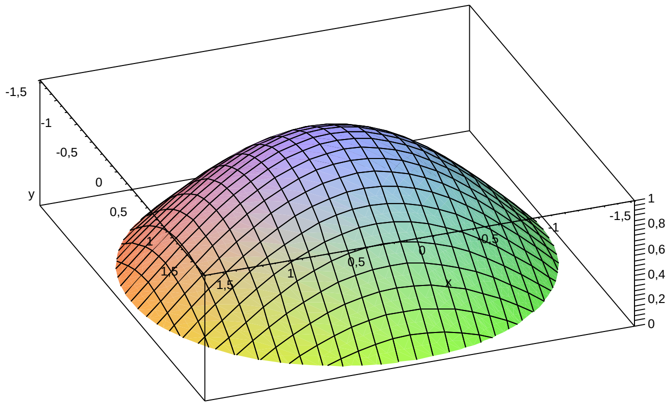
□

8.104. Compute

$$\int_c x^4\ \mathrm{d}x + xy\ \mathrm{d}y,$$

where c is the positively oriented curve going through the vertices $A = [0,0]$; $B = [1,0]$; $C = [0,1]$.

As a truly simple example, we can 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 picture).



First, we determine the Jacobian matrix of the transformation $x = r\cos\theta$, $y = r\sin\theta$

$$D^1G = \begin{pmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{pmatrix}.$$

Hence, the determinant of this matrix is equal to

$$\det D^1G(r,\theta) = r(\sin^2\theta + \cos^2\theta) = r.$$

Therefore, we can calculate directly for the disc S which is the image of the rectangle $(r,\theta) \in [0,R]\times[0,2\pi] = T$. We thus get the area of the disc:

$$\int_S dxdy = \int_0^{2\pi}\int_0^R rdr\,d\theta = \int_0^R 2\pi rdr = \pi R^2.$$

The integration of the function f will be very similar, using multiple integration and integration by parts:

$$\int_S f dxdy = \int_0^{2\pi}\int_0^{\pi/2} r\cos r dr\,d\theta = \pi^2 - 2\pi.$$

8.33. Curve integrals. We often cannot do with integrals over open subsets in $\mathbb{R}^n$ because our quantities are given only on objects which are similar to curves or surfaces in $\mathbb{R}^3$. The previous reasoning about changes of coordinates when computing integrals clarified the intuitive imagination that our process of integration is the sum of volumes of small linearized parallelepipeds multiplied by the value of the integrated function. Extending this idea, we could define integration over such multidimensional surfaces in $\mathbb{R}^n$ directly. However, we will first relieve the integration of dependency on coordinates, and then we will transform it to the well-known integration on $\mathbb{R}^n$.

Recall the calculation of the length of a curve by univariate integrals, which was discussed in paragraph 6.7 on page 353. 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 $\rho : \mathbb{R}^n \to \mathbb{R}$ 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 $L \subset \mathbb{R}^n$.

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:

$$\int_c x^4 \, \mathrm{d}x + xy \, \mathrm{d}y = \iint_D y \, \mathrm{d}x \, \mathrm{d}y = \int_0^1 \int_0^{-x+1} y \, \mathrm{d}x \, \mathrm{d}y =$$
$$= \int_0^1 \left[\frac{y^2}{2}\right]_0^{-x+1} \mathrm{d}x = \int_0^1 \left[\frac{x^2-2x+1}{2}\right] \mathrm{d}x =$$
$$= \frac{1}{2}\left[\frac{x^3}{3} - \frac{2x^2}{2} + x\right]_0^1 = \frac{1}{6}.$$

□

8.105. 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

$$\int_c (xy + x + y) \, \mathrm{d}x + (xy + x - y) \, \mathrm{d}y$$
$$= \iint_D y + 1 - x - 1 \, \mathrm{d}x \, \mathrm{d}y$$
$$= \int_0^1 \int_0^{2\pi} r^2(\sin\varphi - \cos\varphi) \, \mathrm{d}r \, \mathrm{d}\varphi =$$
$$= \int_0^1 r^2 \, \mathrm{d}r \int_0^{2\pi} \sin\varphi - \cos\varphi d\varphi = \frac{1}{3}\big[-\cos\varphi - \sin\varphi\big]_0^{2\pi} = 0.$$

□

8.106. Compute $\int_c (2e^{2x}\sin y - 3y^3)dx + (e^{2x}\cos y + \frac{4}{3}x^3)dy$, where c is the positively oriented ellipse $4x^2 + 9y^2 = 36$.

Solution.: We will use Green's theorem, choosing the linear deformation of polar coordinates $x = 3r\cos\varphi$, $\varphi \in [0, 2\pi]$,

Now, we will proceed in a much similar way. Let us consider a differentiable curve $c(t)$ in $\mathbb{R}^n$, $t \in [a, b]$ and assume that a differentiable function f is defined on a neighborhood of its values. The differential of this function gives, for every tangent vector, the increase of the function in the given direction. It is expressed by the differential of the composite mapping $f \circ c$ by

$$d(f \circ c)(t) = \frac{\partial f}{\partial x_1}(c(t))c_1'(t) + \cdots + \frac{\partial f}{\partial x_n}(c(t))c_n'(t).$$

We can thus try to define the value of the integral in this way (we keep writing df symbolically to emphasize which object is integrated, just like we wrote dx in the case of univariate integrals)

$$\int_M f \, d\mathrm{vol}_M = \int_a^b \left(\frac{\partial f}{\partial x_1}(c(t))c_1'(t) + \cdots + \frac{\partial f}{\partial x_n}(c(t))c_n'(t)\right)dt,$$

and we can 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

$$\int_{\tilde{a}}^{\tilde{b}} \left(\frac{\partial f}{\partial x_1}(c(\psi(s)))c_1'(\psi(s)) + \dots \right.$$
$$\left. + \frac{\partial f}{\partial x_n}(c(\psi(s)))c_n'(\psi(s))\right)\frac{d\psi}{ds} \, ds,$$

and the theorem about coordinate transformations for univariate integrals gives just the same value if we have $\frac{d\psi}{ds} > 0$, i. e., if we keep the orientation of the curve, and the same value up to sign if the derivative of the transformation is negative.

Precisely speaking, we have learned to integrate the differential df of a function over curves. However, it may be the case that the connection with integration of functions is not apparent. We clearly cannot get the length of the curve if we select a constant function with value one for f. We need a geometric point of view to explain this. The size of a vector is given by a quadratic form, rather than a linear one. However, if we take the square root of the values of a (positively definite) quadratic form, we get a linear form (up to sign, see above). We will get back to these connections shortly.

8.34. Vector fields and linear forms. In the previous paragraph, the parametrization of a curve was used to obtain a *tangent vector* $c'(t) \in \mathbb{R}^n$ for every point in the image M of the curve. We thus have a mapping $X : M \to M \times \mathbb{R}^n$, $c(t) \mapsto (c(t), c'(t))$. We talk about the *vector field X along the curve M*.

In general, we define a *vector field* X on an open set $U \subset \mathbb{R}^n$ as assigning the vector $X(x) \in \mathbb{R}^n$ in the direction space of the Euclidean space $\mathbb{R}^n$ to every of its points x in the considered domain.

If a vector field X on an open set $U \subset \mathbb{R}^n$ is given, then we can define for every differentiable function f on U its derivative in the direction of the vector field X in terms of the directional derivative by the formula

$$X(f) : U \to \mathbb{R}, \qquad X(f)(x) = d_{X(x)}f.$$

Therefore, if we have, in coordinates, $X(x) = (X_1(x), \dots X_n(x))$, then

$$X(f)(x) = X_1(x)\frac{\partial f}{\partial x_1}(x) + \cdots + X_n(x)\frac{\partial f}{\partial x_n}(x).$$

The simplest vector fields will have all coordinate functions equal to zero except for one function X_i which will be constantly equal

$y = 2r\sin\varphi \qquad\qquad r \in [0,1],$
leading to (the Jacobian of the transformation is $6r$):

$$\int\limits_c (2e^{2x}\sin y - 3y^3)dx + (e^{2x}\cos y + \frac{4}{3}x^3)\,\mathrm{d}y =$$

$$= \iint\limits_D 2e^{2x}\cos y + 4x^2 - (2e^{2x}\cos y - 9y^2)\ \mathrm{d}x\ \mathrm{d}y =$$

$$= \int\limits_0^1\int\limits_0^{2\pi} 6r\left[4(3r\cos\varphi)^2 + 9(2r\sin\varphi)^2\right] =$$

$$= 216\int\limits_0^1 r^3\ \mathrm{d}r\int\limits_0^{2\pi}\mathrm{d}\varphi = 216\cdot\big[\frac{r^4}{4}\big]_0^1\cdot 2\pi = 108\pi.$$

□

8.107. Compute

$$\int\limits_c (e^x \ln y - y^2x)dx + \left(\frac{e^x}{y} - \frac{1}{2}x^2y\right)\mathrm{d}y,$$

where c is the positively oriented circle $(x-2)^2 + (y-2)^2 = 1$.

Solution.

$$\int\limits_c (e^x\ln y - y^2x)dx + (\frac{e^x}{y} - \frac{1}{2}x^2y)dy =$$

$$= \iint\limits_D \frac{e^x}{y} - xy - \frac{e^x}{y} + 2xy\ \mathrm{d}x\ \mathrm{d}y =$$

$$= \int\limits_0^1\int\limits_0^{2\pi} r(r\cos\varphi + 2)\cdot(r\sin\varphi + 2)\ \mathrm{d}r\ \mathrm{d}\varphi =$$

$$= \int\limits_0^1\int\limits_0^{2\pi} r^3\sin\varphi\cos\varphi + 2r^2(\sin\varphi + \cos\varphi) + 4r\ \mathrm{d}r\ \mathrm{d}\varphi =$$

$$= \frac{1}{4}\int\limits_0^{2\pi}\sin\varphi\cos\varphi\ \mathrm{d}\varphi + \frac{2}{3}\int\limits_0^{2\pi}\sin\varphi + \cos\varphi\ \mathrm{d}\varphi + 4\pi =$$

$$= \frac{1}{3}\big[\frac{\sin^2\varphi}{2}\big]_0^{2\pi} + \big[-\cos\varphi + \sin\varphi\big]_0^{2\pi} + 4\pi = 4\pi.$$

□

8.108. Calculate the integral

$$\int\limits_c (e^x\sin y - xy^2)dx + \left(e^x\cos y - \frac{1}{2}x^2y\right)\mathrm{d}y,$$

where c is the positively oriented circle $x^2 + y^2 + 4x + 4y + 7 = 0$.
○

8.109. Compute

$$\int\limits_c (3y - e^{\sin x})\ \mathrm{d}x + (7x + \sqrt{y^4 + 1})\ \mathrm{d}y,$$

to one. Such a field then corresponds to the partial derivative with respect to the variable x_i. This is also matched by the common notation

$$X(x) = X_1(x)\frac{\partial}{\partial x_1} + \cdots + X_n(x)\frac{\partial}{\partial x_n}.$$

The set of all possible tangent vectors at the points of an open subset $U \in \mathbb{R}^n$ is called the *tangent space* TU. The vector space of all vectors at a point x is denoted by T_xU. We use the notation $\mathcal{X}(U)$ for the set of all smooth vector fields on U. 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.

When we studied the vector spaces, we found out as early as in the second chapter that we need the so-called linear forms. They were defined in paragraph 2.39 on page 103. The same idea suggests itself now. The linear form on the direction space $\mathbb{R}^n$ of our Euclidean space $\mathbb{R}^n$ assigned to a point $x \in \mathbb{R}^n$ is a linear mapping defined on the tangent space T_xU. Having a mapping $\eta : U \subset \mathbb{R}^n \to \mathbb{R}^{n*}$ on an open subset U, we talk about a *linear form η on U*.

Every differentiable function f on an open subset $U \subset \mathbb{R}^n$ defines a linear form df on U. We use the notation $\Omega^1(U)$ for the set of all smooth linear forms on U.

It is apparent that in the coordinates $(x_1,\dots,x_n)$, 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 is evaluated at a vector field $X(x) = X_1(x)\frac{\partial}{\partial x_1} + \cdots + X_n(x)\frac{\partial}{\partial x_n}$ by

$$\eta(X(x)) = \eta_1(x)X_1(x) + \cdots + \eta_n(x)X_n(x).$$

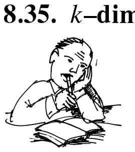
If the form η is the differential of a function f, we get just the expression $X(f)(x) = df(X(x))$ used above.

Let us notice that we have actually defined the integral of any linear form over (non-parametrized) curves M in terms of an arbitrary parametrization $c(t)$

$$\int_M \eta = \int_a^b \eta(c(t))(c'(t))\,dt,$$

since although we worked with the function differential back then, we actually verified that the value of the integral was independent of the choice of parametrization for any linear form.

We can also notice that we need not write any symbol denoting which concept of a volume we are integrating with respect to. It is given by the definition of a linear form.



8.35. k–dimensional surfaces and k–forms. Instead of parametrized curves, we will now work with differentiable mappings $\varphi : V \subset \mathbb{R}^k \to \mathbb{R}^n$, $k \le n$, with injective differential $d\varphi(u)$ at every point of its open domain V. Such mappings are called *immersions*.

A subset $M \subset \mathbb{R}^n$ is called a manifold of dimension r iff every point $x \in M$ has a neighborhood U which is the image of such an immersion $\varphi : V \subset \mathbb{R}^k \to M \subset \mathbb{R}^n$ which can be extended to a mapping $\tilde{\varphi} : V \times \tilde{V} \to \mathbb{R}^n$ which is a diffeomorphism and $\tilde{\varphi}^{-1}(M) = V \times \{0\}$. This definition, which might seem complicated at the first sight, is illustrated by a picture. Manifolds can be typically given by implicit mappings, see paragraph 8.18 and the discussion in 8.19.

where c is the positively oriented circle $x^2 + y^2 = 9$. ○

8.110. Compute the integral

$$\int_c (\frac{1}{x} + 2xy - \frac{y^3}{3})\,dx + (\frac{1}{y} + x^2 + \frac{x^3}{3})\ \mathrm{d}y,$$

where c is the positively oriented boundary of the set $D = \{(x, y) \in \mathbb{R}^2 : 4 \le x^2 + y^2 \le 9, \frac{x}{\sqrt{3}} \le y \le \sqrt{3}x\}$. ○

8.111. Remark. An important corollary of *Green's theorem* is the formula for computing the area D that is bounded by a curve c.

$$m(D) = \frac{1}{2}\int_c -y\ \mathrm{d}x + x\ \mathrm{d}y.$$

8.112. Compute the area given by the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Solution. Using the formula ||8.111|| and the transformation $x = a\cos t$, $y = b\sin t$, we get for $t \in [0, 2\pi]$ that

$$\begin{aligned} m(D) &= \frac{1}{2}\int_c -y\ \mathrm{d}x + x\ \mathrm{d}y = \\ &= \frac{1}{2}\int_0^{2\pi} a\cos t \cdot b\cos t dt - \frac{1}{2}\int_0^{2\pi} b\sin t \cdot (-a\sin t)dt = \\ &= \frac{1}{2}ab\int_0^{2\pi} \cos^2 t dt + \frac{1}{2}ab\int_0^{2\pi} \sin^2 t dt = \\ &= \frac{1}{2}ab\int_0^{2\pi} \cos^2 t + \sin^2 t dt = \frac{1}{2}ab2\pi = \pi ab, \end{aligned}$$

which is indeed the well-known formula for the area of an ellipse with semi-axes a and b.

□

8.113. Find the area bounded by the cycloid which is given parametrically as $\psi(t) = [a(t - \sin t); a(1 - \cos t)]$, for $a \ge 0$, $t \in (0, 2\pi)$, 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

$$m(D) = \tfrac{1}{2}\int_{c_1} -y\ \mathrm{d}x + x\ \mathrm{d}y + \tfrac{1}{2}\int_{c_2} -y\ \mathrm{d}x + x\ \mathrm{d}y.$$

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

$$\frac{1}{2}\int_{C_1} -y\ \mathrm{d}x + x\ \mathrm{d}y = \frac{1}{2}\int_0^{2a\pi} 0 \cdot 1\,\mathrm{d}t + \int_0^{2a\pi} t \cdot 0 dt = 0.$$

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 mapping φ from the definition is called a *local parametrization of the manifold M*. The tangent space TM to the manifold M is the collection of vector subspaces $T_xM \subset T_x\mathbb{R}^n$ which contain all vectors which are tangent to the curves in M. Clearly, every parametrization φ defines a diffeomorphism

$$\varphi_* : TV \to T\varphi(V) \subset TM, \varphi_*(c'(t)) = \frac{d}{dt}\varphi(c(t)).$$

This definition is independent of the choice of the curve representing the vector $c'(t)$ because we need to know only the first derivatives when calculating the right-hand derivatives.

Our definitions are a straightforward generalization of the concepts of a curve and a surface in the plane or the space and of their tangent lines or planes. 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 numeral 8 parametrized with a mapping φ with everywhere-injective differential. However, we will be unable to satisfy the second property in a neighborhood of the point where the two branches of the curve meet.

To be able to talk about a volume on k-dimensional manifolds in linear approximations, as we did with curves, we need objects which will be linear at every point in k distinct vector arguments and will assign a number to them. Moreover, we will require that interchanging any pair of arguments swap the sign, in accordance with the orientations. We have already met such objects, namely in paragraph 2.44 on page 107, and mainly when calculating the volumes of parallelepipeds using determinants in paragraph 4.22 on page 223.

Exterior differential forms

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 $\Lambda^k(T_x\mathbb{R}^n)^*$. In short, we talk about an exterior k–form at the point x.

The assignment of a k–form $\eta(x)$ 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 $\Omega^k(U)$.

Now, let us consider a smooth parametrization $\varphi : V \to M$ of a manifold M, an $\eta(\varphi(u)) \in \Lambda^k(T_{\varphi(u)}\mathbb{R}^n)$, and choose arbitrarily k vectors $X_1(u), \dots, X_k(u)$ in the tangent space T_uV. Just like in the case of linear forms, we can evaluate the form η at the images of the vectors X_i using the parametrization φ. This operation is called the *pullback of the form η by φ*.

$$\begin{aligned} &\varphi^*(\eta(\varphi(u)))(X_1(u), \dots, X_k(u)) \\ &\qquad = \eta(\varphi(u))(\varphi_*(X_1(u)), \dots, \varphi_*(X_1(u))). \end{aligned}$$

We can compute directly from the definition that, for instance,

$$\varphi^*(dx_i)(\frac{\partial}{\partial u_k}) = dx_i\big(\varphi_*(\frac{\partial}{\partial u_k})\big) = \frac{\partial \varphi_i}{\partial u_k},$$

and so

$$\varphi^*(dx_i) = \frac{\partial \varphi_i}{\partial u_1}du_1 + \dots + \frac{\partial \varphi_i}{\partial u_n}du_n. \tag{8.1}$$

Another immediate consequence of the definition is the relation for pullbacks of an arbitrary k–form by composing two diffeomorphisms. Verify the following equality by yourselves!

$$(\varphi \circ \psi)^*\alpha = \psi^*\big(\varphi^*\alpha\big). \tag{8.2}$$

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

$$\frac{1}{2}\int\limits_{c_2} -y\ \mathrm{d}x + x\ \mathrm{d}y = \frac{1}{2}\int\limits_{2\pi}^{0} a(t-\sin t)\cdot a(\sin t)\ \mathrm{d}t -$$

$$-\frac{1}{2}\int\limits_{2\pi}^{0} a(1-\cos t)\cdot a(1-\cos t)\ \mathrm{d}t =$$

$$=\frac{1}{2}a^2\int\limits_{0}^{2\pi} t\sin t - \sin^2 t - 1 + 2\cos t - \cos^2 t\ \mathrm{d}t =$$

$$=\frac{1}{2}a^2\int\limits_{2\pi}^{0} t\sin t + 2\cos t - 2\ \mathrm{d}t =$$

$$=\frac{1}{2}a^2[-t\cos t - \sin t + 2\cos t - 2]_{2\pi}^{0} = 3\pi a^2.$$

□

J. Applications of Stoke's theorem – the Gauss–Ostrogradsky theorem

8.114. Compute $I = \iint\limits_S x^3\ \mathrm{d}y\ \mathrm{d}z + y^3\ \mathrm{d}x\ \mathrm{d}z + z^3\ \mathrm{d}x\ \mathrm{d}y$,where S is given by the sphere $x^2 + y^2 + z^2 = 1$.

Solution. It is advantageous to work in spherical coordinates

$$\begin{aligned} x &= \rho\sin\varphi\cos\psi & \rho &= [0,1],\\ y &= \rho\sin\varphi\sin\psi & \varphi &= [0,\pi],\\ z &= \rho\cos\varphi & \psi &= [0,2\pi]. \end{aligned}$$

The Jacobian of this transformation is $-\rho^2\sin\varphi$.

The given integral is then equal to

$$I = \iint\limits_S x^3\ \mathrm{d}y\ \mathrm{d}z + y^3\ \mathrm{d}x\ \mathrm{d}z + z^3\ \mathrm{d}x\ \mathrm{d}y =$$

$$= \iiint\limits_V 3x^2 + 3y^2 + 3z^2\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z\ =$$

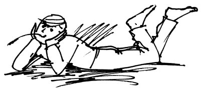
$$= 3\int\limits_0^1\int\limits_0^{2\pi}\int\limits_0^{\pi} \rho^2\sin\varphi\left(\rho^2\sin^2\varphi\cos^2\psi + \rho^2\sin^2\varphi\sin^2\psi +\right.$$

$$\left. +\rho^2\cos^2\varphi\right)\ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}\psi =$$

$$= 3\int\limits_0^1\int\limits_0^{2\pi}\int\limits_0^{\pi} \rho^4\sin\varphi\left(\sin^2\varphi\left(\cos^2\psi + \sin^2\psi\right) +\right.$$

$$\left. + \cos^2\varphi\right)\ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}\psi =$$

A smooth k–form η on a k–dimensional submanifold is such a mapping $M \to \Lambda^k(T_xM)^*$ that the pullback of this form by any parametrization yields a smooth exterior k–form on V. We will use the notation $\Omega^k(M)$ for the set of all smooth exterior k–forms on M.



8.36. Outer 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:

$$(\alpha\wedge\beta)(X_1,\dots,X_{k+\ell}) =$$
$$\frac{1}{k!\ell!}\sum_{\sigma\in\Sigma_{k+\ell}} \operatorname{sign}(\sigma)\alpha(X_{\sigma(1)},X_{\sigma(k)})\beta(X_{\sigma(k+1)},X_{\sigma(k+\ell)}).$$

It is clear from the definition that $\alpha\wedge\beta$ is indeed a $(k+\ell)$–form. In the simplest case of 1–forms, the definitions says that

$$(\alpha\wedge\beta)(X,Y) = \big(\alpha(X)\beta(Y) - \alpha(Y)\beta(X)\big).$$

In the case of a 1–form α and a k–form β, we get

$$(\alpha\wedge\beta)(X_0,X_1,\dots,X_k) =$$
$$\sum_{j=0}^{k}(-1)^j\alpha(X_j)\beta(X_0,\dots,\hat{X}_j,\dots,X_k),$$

where that hat indicates omission of the corresponding argument. The outer product of finitely many form is defined analogously (either directly by a similar formula, or we can notice that the outer product of two forms is associative – think this out by yourselves!).

We will use the same notation for forms in $\Omega^k(M)$. Just as we had generators $\frac{\partial}{\partial x_i}$ of all vector spaces in $\mathcal{X}(\mathbb{R}^n)$, all linear forms in $\Omega^1(\mathbb{R}^n)$ are generated by the forms dx_i. Their outer products

$$\varepsilon_{i_1\dots i_k} = dx_{i_1}\wedge\cdots\wedge dx_{i_k}$$

with k-tuples of indeces $i_1 < i_2 < \cdots < i_k$ generate the whole space $\Omega^k(\mathbb{R}^n)$. 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 $\alpha_{i_1\dots i_k}(x)$

$$\alpha(x) = \sum_{i_1<\dots<i_k}\alpha_{i_1\dots i_k}(x)dx_{i_1}\wedge\cdots\wedge dx_{i_k}.$$

We can also notice that 0–forms $\Omega^0(\mathbb{R}^n)$ are, by the definition, functions on $\mathbb{R}^n$. The outer product of a 0–form f and a k–form α is just the multiple of the form α by the function f.

Verify that the pullback of the outer product by a diffeomorphism $\varphi : V \to U$ satisfies

$$\varphi^*(\alpha\wedge\beta) = \varphi^*\alpha\wedge\varphi^*\beta.$$

8.37. Integration of exterior forms on $\mathbb{R}^n$. Exterior n–forms on $\mathbb{R}^n$ have a unique behavior as we have a single non-decreasing sequence of n indeces $i_k = k$ at our disposal. Therefore, the whole space $\Omega^n(\mathbb{R}^n)$ is generated by a single form $\varepsilon_{12\dots n}$ and it can be identified with the space of functions $f(x)$. Each of these functions defines, at every point $x\in\mathbb{R}^n$, the infinitesimal calculation of the volume by the n–form

$$\omega(x) = f(x)dx_1\wedge\cdots\wedge dx_n, \tag{8.3}$$

$$= 3\int_0^1\int_0^{2\pi}\int_0^{\pi} \rho^4 \sin\varphi \,\mathrm{d}\rho\,\mathrm{d}\varphi\,\mathrm{d}\psi = 3\cdot\left[\frac{\rho^5}{5}\right]_0^1 [\psi]_0^{2\pi}[\cos\varphi]_0^{\pi} =$$

$$= 3\cdot\frac{1}{5}\cdot 2\pi\cdot[-1-1] = -\frac{12}{5}\pi.$$

□

8.115. The vector form of the Gauss–Ostrogradsky theorem. The divergence of a vector field $F(x,y,z) = f(x,y,z)\frac{\partial}{\partial x} + g(x,y,z)\frac{\partial}{\partial y} + h(x,y,z)\frac{\partial}{\partial z}$ is defined as $\operatorname{div} X := f_x + g_y + h_z$. Then, the Gauss–Ostrogradsky theorem can be formulated as follows:

$$\iiint_V \operatorname{div}\vec{F}(x,y,z)\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z\ = \iint_S \vec{F}(x,y,z)\cdot\vec{n}(x,y,z)\ \mathrm{d}S,$$

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).

8.116. Find the flow of the vector field given by the function $F = (xy^2, yz, x^2z)$ over the cylinder $x^2+y^2 = 4, z = 1, z = 3$.

Solution. First of all, we compute the divergence of the vector field:

$$\operatorname{div} F = \nabla\cdot F = (\frac{\partial(xy^2)}{\partial x} + \frac{\partial(yz)}{\partial y} + \frac{\partial(x^2z)}{\partial z}) = y^2 + z\ + x^2.$$

Therefore, the flow T of the vector filed is equal to

$$\iiint_V y^2 + z\ + x^2\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z\ =$$

$$= \int_0^2\int_0^{2\pi}\int_1^3 \rho\cdot(\rho^2\sin^2\varphi + z\ + \rho^2\cos^2\varphi)\ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}z\ =$$

$$= \int_0^2\int_0^{2\pi}\int_1^3 \rho\cdot(\rho^2(\sin^2\varphi + \cos^2\varphi) + z\ )\ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}z\ =$$

$$= \int_0^2\int_0^{2\pi}\int_1^3 \rho\cdot(\rho^2(\sin^2\varphi + \cos^2\varphi) + z\ )\ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}z\ =$$

$$= \int_0^2\int_0^{2\pi}\int_1^3 \rho^3 + \rho z\ \ \mathrm{d}\rho\ \mathrm{d}\varphi\ \mathrm{d}z\ =$$

$$= 2\pi\int_0^2\int_1^3 \rho^3 + \rho z\ \ \mathrm{d}\rho\ \mathrm{d}z\ = 2\pi\int_1^3 [\frac{\rho^4}{4} + \frac{\rho^2}{2}z]_0^2\ \mathrm{d}z\ =$$

$$= 2\pi\int_1^3 4 + 2z\ \mathrm{d}z\ = 2\pi[4z + z^2]_1^3 = 2\pi[12 + 9 - 4 - 1] = 32\pi.$$

□

i. e., f gives, at every point, the scale with which the standard volume is to be taken. We have thus obtain a new interpretation of our earlier procedure of integration of functions f on Riemann-measurable open subsets $U\subset\mathbb{R}^n$.

First, we define a form $\omega_{\mathbb{R}^n}$, giving the standard n–dimensional volume of parallelograms, i. e., in the standard coordinates, we will have

$$\omega_{\mathbb{R}^n} = dx_1\wedge\cdots\wedge dx_n.$$

If we want to integrate a function $f(x)$ "in the old fashion", we consider the form $\omega = f\omega_{\mathbb{R}^n}$ instead, i. e. ω will have the form (8.3) in the standard coordinates. We define

$$\int_U \omega = \int_U f(x)dx_1\wedge\cdots\wedge dx_n = \int_U f(x)\,dx_1\dots dx_n,$$

where there is the Riemann integral of a function on the right-hand side. We can notice, that the n-form on the left-hand side is independent of the choice of coordinates.

If we want to express the form ω in different coordinates using a diffeomorphism $\varphi : V\to U$, it means we will evaluate ω at a point $\varphi(u) = x$ at the values of the vectors $\varphi_*(X_1),\dots,\varphi_*(X_n)$. However, this means we will integrate the form $\varphi^*\omega$ in coordinates $(u_1,\dots,u_n)$, and we can easily compute (look at (8.1) in paragraph 8.35) that

$$\begin{aligned}(\varphi^*\omega)(u) &= f(\varphi(u))\big(\frac{\partial\varphi_1}{\partial u_1}du_1 + \cdots + \frac{\partial\varphi_1}{\partial u_n}du_n\big)\wedge\dots\\ &\qquad\wedge\big(\frac{\partial\varphi_n}{\partial u_1}du_1 + \cdots + \frac{\partial\varphi_n}{\partial u_n}du_n\big)\\ &= f(\varphi(u))\det(D^1\varphi(u))du_1\wedge\cdots\wedge du_n.\end{aligned}$$

Substituting into our interpretation of the integral, we get

$$\int\varphi^*(f\omega) = \int_V f(u)\det(D^1\varphi(u))du_1\cdots du_n,$$

which is, by the theorem on transformation of variables from paragraph 8.31, the same value if the determinant of the Jacobian matrix keeps being positive, and the same value up to sign if it is negative.

Our new interpretation thus yields a geometrical sense for the integral of an n–form on $\mathbb{R}^n$, supposing the corresponding Riemann-integral exists in some (hence any) coordinates. This integration takes into account the orientation of the area we are integrating over.

8.38. Integration of exterior forms on manifolds. Now, we are almost ready for the definition of an integral of a k–formy on a k–dimensional oriented manifold. For the sake of simplicity, we will examine smooth forms ω with compact support.

First, let us assume that we are given a k–dimensional manifold $M\subset\mathbb{R}^n$ and one of its local parametrizations $\varphi : V\subset\mathbb{R}^k\to U\subset M\subset\mathbb{R}^n$. 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 which differ by diffeomorphisms with positive determinants of their Jacobian matrices. The orientation will be the other one in the case of negative determinants. Therefore, we apparently have exactly two orientations on every connected manifold. Fixing either of them, we thereby restrict the set of parametrizations which are compatible with this orientation.

8.117. 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

$$\iiint\limits_V 2z \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z = \int\limits_0^2 \int\limits_0^\pi \int\limits_0^{2\pi} \rho^2 \sin\varphi \cdot 2\rho \cos\varphi \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\psi =$$

$$= 2 \int\limits_0^2 \rho^3 \, \mathrm{d}\rho \int\limits_0^{2\pi} \mathrm{d}\psi \int\limits_0^\pi \sin\varphi \cos\varphi \, \mathrm{d}\varphi =$$

$$= 2[\frac{\rho^4}{4}]_0^2 \cdot [\psi]_0^{2\pi} \cdot [\frac{\sin^2\varphi}{2}]_0^\pi =$$

$$= 2 \cdot \frac{16}{4} \cdot 2\pi \cdot 0 = 0.$$

□

K. First-order differential equations

8.118. Find all solutions of the differential equation

$$y' = \tfrac{\sqrt{1-y^2}}{\cos^2 x} \left(1 + \cos^2 x\right).$$

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

$$\tfrac{1}{\sqrt{1-y^2}} \, dy = \tfrac{1+\cos^2 x}{\cos^2 x} \, dx.$$

Since

$$\int \tfrac{1+\cos^2 x}{\cos^2 x} \, dx = \int \tfrac{1}{\cos^2 x} + 1 \, dx,$$

we can integrate using the basic formulae, thereby obtaining

$$\arcsin y = \tan x + x + C, \qquad C \in \mathbb{R}. \tag{8.6}$$

However, we must keep in mind that the division by the expression $\sqrt{1 - y^2}$ is valid only if it is non-zero, i. e., only for $y \neq \pm 1$. Substituting the constant functions $y \equiv 1$, $y \equiv -1$ into the given differential equation, we can immediately see that they satisfy it. We have thus obtained two more solutions, which are called singular. We do not have to pay attention to the case $\cos x = 0$ since this only loses points of the domains (but not any solutions).

Now, we will comment on several parts of the computation. The expression $y' = dy/dx$ allows us to make many symbolic manipulations. For instance, we have

$$\tfrac{dz}{dy} \cdot \tfrac{dy}{dx} = \tfrac{dz}{dx}, \qquad \tfrac{1}{\frac{dy}{dx}} = \tfrac{dx}{dy}.$$

From now on, we will always proceed in this fashion, and we will talk about *oriented manifolds*.

Now, let us select a form ω with compact support inside the image of a parametrization $U \subset M$ of an oriented manifold M. The 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 using the same definition. Let us denote

$$\varphi^*(\omega)(u) = f(u) du_1 \wedge \cdots \wedge du_k.$$

Invoking the relation (8.2) for the pullback of a form by a composite mapping, we get

$$\begin{aligned} \int_M \omega = \int_{\mathbb{R}^k} \tilde{\varphi}^*(\omega) &= \int_{\mathbb{R}^k} \psi^*(\varphi^* \omega) \\ &= \int_{\mathbb{R}^k} \psi^*\big(f \, du_1 \wedge \cdots \wedge du_k\big) \\ &= \int_{\mathbb{R}^k} f(\psi(v)) \det(D^1\psi)(v) dv_1 \cdots dv_k. \end{aligned}$$

This is again the same value as $\int_{\mathbb{R}^k} \varphi^* \omega$.

This proves the correctness of our definition of the integral $\int_M \omega$ provided the integrated k–form has compact support lying in the image of a single parametrization.

However, typical manifolds M are given by implicit equations; e. g. $x^2 + y^2 + z^2 = 1$ defines the surface of the unit ball, i. e. the sphere $S^2 \subset R^3$. 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 areas, and we integrate on each of them separately. However, we will mention a global definition which is more advantageous from the technical point of view.

8.39. Unit decomposition. 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 can 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*.

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 (||8.6||) in the suggesting form

$$(8.7)\qquad y = \sin(\tan x + x + C)\,, \qquad C \in \mathbb{R}.$$

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 (||8.7||) 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 (||8.6||). 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. □

8.119. Find the general solution of the equation $y' = (2-y)\tan x$.

Solution. Again, we are given a differential equation with separated variables.

We have

$$\frac{dy}{dx} = (2-y)\tan x,$$
$$-\frac{dy}{y-2} = \frac{\sin x}{\cos x}\,dx,$$
$$-\ln|\,y-2\,| = -\ln|\,\cos x\,| - \ln|\,C\,|, \qquad C \neq 0.$$

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

$$\ln|\,y-2\,| = \ln|\,C\cos x\,|, \qquad C \neq 0,$$
$$|\,y-2\,| = |\,C\cos x\,|, \qquad C \neq 0,$$
$$y-2 = C\cos x, \qquad C \neq 0,$$

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

$$y = 2 + C\cos x, \qquad C \in \mathbb{R}.$$ □

8.120. 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

$$y' = f(x)g(y), \qquad y(x_0) = y_0$$

Now, recall the smooth variants of indicator functions from paragraph 6.7. For every pair of positive numbers $\varepsilon < r$, we constructed a function $f_{\varepsilon,r}(t)$ 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 $f(t) \neq 0$ if and only if $|t| \le r + \varepsilon$.

Now, if we define

$$\chi_{r,\varepsilon,x_0}(x) = f_{\varepsilon,r}(|x - x_0|),$$

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.

Lemma (Whitney's theorem)**.** *Every closed set $K \subset \mathbb{R}^n$ is the set of all zero points of some smooth function.*

Proof. The idea of the proof is quite simple. If $K = \mathbb{R}^n$, the zero function is convenient, so we can further assume that $K \neq \mathbb{R}^n$.

An 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 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, not only 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 $\mathbb{R}^n$, but 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 $\frac{\partial^r f}{\partial x_{i_1}\cdots\partial x_{ir}}$.

It is apparent from the definition that the function $f(x)$ satisfies the conditions of the lemma. □

Unit decomposition on a manifold

Theorem. *Consider a manifold $M \subset \mathbb{R}^n$ and one of its locally finite covers by open images U_i of parametrizations φ_i. Then, there exists a system of smooth 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 *unit decomposition subordinate to a locally finite cover* of a manifold by parametrizations.

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:

$$\begin{aligned}\left(1+\mathrm{e}^x\right) y\,dy/dx &= \mathrm{e}^x,\\ y\,dy &= \frac{\mathrm{e}^x}{1+\mathrm{e}^x}\,dx,\\ \frac{y^2}{2} = \ln\left(1+\mathrm{e}^x\right) + \ln|\,C\,|&, \quad C \neq 0,\\ \frac{y^2}{2} = \ln\left(C\left[1+\mathrm{e}^x\right]\right)&, \quad C > 0.\end{aligned}$$

The substitution $y = 1$, $x = 0$ then gives

$$\tfrac{1}{2} = \ln\left(C \cdot 2\right), \quad \text{i. e.} \quad C = \tfrac{\sqrt{\mathrm{e}}}{2}.$$

We have thus found the solution

$$\tfrac{y^2}{2} = \ln\left(\tfrac{\sqrt{\mathrm{e}}}{2}\left[1+\mathrm{e}^x\right]\right),$$

i. e.,

$$y = \sqrt{2\ln\left(\tfrac{\sqrt{\mathrm{e}}}{2}\left[1+\mathrm{e}^x\right]\right)}$$

on a neighborhood of the point [0, 1] where $y > 0$. □

8.121. 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

$$\begin{aligned}\frac{dy}{y^2+1} &= \frac{dx}{x+1},\\ \arctan y = \ln|\,x+1\,| + C&, \quad C \in \mathbb{R}.\end{aligned}$$

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

$$y(x) = \tan\left(\ln|\,x+1\,| + \tfrac{\pi}{4}\right)$$

on a neighborhood of the point [0, 1]. □

8.122. Solve

$$y' = \frac{x+y+1}{2x+2y-1}. \tag{8.8}$$

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

$$\begin{vmatrix} f(x,y) & f'_y(x,y)\\ f'_x(x,y) & f''_{xy}(x,y)\end{vmatrix} = 0, \quad x \in (a,b),\ y \in (c,d).$$

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 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, having been just proved. 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 non-zero for all $x \in M$, so we can consider not the functions $g_i(x)$ restricted to M, but rather the functions $f_i(x) = g_i(x)/g(x)$, which already have both of the required properties of the theorem. □

8.40. Integration of k–forms on manifolds. Finally, we are ready for the definition of integral of k–forms on k–dimensional manifolds. Let us thus consider a manifold $M \subset \mathbb{R}^n$ and a form $\omega \in \Omega^k(M)$ having 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 unit decomposition f_i subordinate 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. Actually, we can assume that our sum is finite, since it suffices to consider 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 our choices.

To this purpose, let us choose another parametrization $\varphi : V \to U$ of a piece U of the manifold M, and let us have a look at the contribution of the integral over U. We get

$$\int_U \omega = \sum_i \int_{V_i \cap V} (f_i \circ \varphi)(\varphi^*\omega) = \int_V \varphi^*\omega.$$

Therefore, if we select a different cover and unit decomposition, we can do the above reasoning for a common refinement of these covers and verify that the expression we have defined is actually independent of all of our choices (think this out in detail!).

8.41. Exterior differential of exterior forms. 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 can be extended to functions on manifolds M, where the differential is a linear form on M. The following theorem extends this differential to arbitrary exterior forms on manifolds $M \subset \mathbb{R}^n$.

Exterior differential

Theorem. *There is a unique mapping $d : \Omega^k(M) \to \Omega^{k+1}M$, for all manifolds $M \subset \mathbb{R}^n$ and $k = 0, \dots, k$, such that*

- *d is linear with respect to multiplication by real numbers*
- *for $k = 0$, it is a differential of functions*

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 ($\|8.8\|$) yields

$$\begin{aligned} z' - 1 &= \frac{z\ +1}{2z-1}, \\ \frac{dz}{dx} &= \frac{z\ +1}{2z-1} + 1, \\ \frac{dz}{dx} &= \frac{3z}{2z-1}, \\ \left(\frac{2}{3} - \frac{1}{3z}\right) dz &= 1\, dx, \\ \frac{2}{3}z - \frac{1}{3}\ln|\, z\,| &= x + C, \quad C \in \mathbb{R}, \end{aligned}$$

or

$$\tfrac{2}{3}z - \tfrac{1}{3}\ln|\, Cz\,| = x, \quad C \neq 0.$$

Now, we must get back to the original variable y in one of these forms. The general solution can be written as

$$\tfrac{2}{3}x + \tfrac{2}{3}y - \tfrac{1}{3}\ln|\, x + y\,| = x + C, \quad C \in \mathbb{R},$$

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 $3z$). □

8.123. 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

$$y' = \tfrac{y}{x} \ln \tfrac{y}{x}.$$

- *$d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k \alpha \wedge (d\beta)$, where $\alpha \in \Omega^k(M)$*
- *$d(df) = 0$ for every function f on M.*

The mapping d is called an *exterior differential.*

Proof. The k–form can be written locally in the form

$$\alpha = \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} dx_{i_1} \wedge \dots \wedge dx_{i_k}.$$

If the differential d exists, then by the required properties, it must be equal to

$$\begin{aligned} d\alpha &= \sum_{i_1 < \dots < i_k} da_{i_1 \dots i_k} dx_{i_1} \wedge \dots \wedge dx_{i_k} \\ &= \sum_{i_1 < \dots < i_k} \frac{\partial a_{i_1 \dots i_k}}{\partial x_i} dx^i \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k}. \end{aligned}$$

Indeed, the basis linear forms dx_i 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 $d(f\beta) = df \wedge \beta + f\, d\beta$. (Think out the details!).

On the other hand, if we define the differential d in coordinates this way, we can easily verify all of the required properties. (Finish by yourselves!) □

8.42. Manifolds with boundary. In practical problems, we often work with manifolds M like an open ball in the three-dimensional space, for example. 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 a curve M which is the image of a parametrization $\varphi : [a, b]$, then we get directly from the definition that

$$\int_M df = \int_a^b d(f \circ \varphi)(t)\, dt = f(\varphi(b)) - f(\varphi(a)).$$

Therefore, the result is independent of not only the selected parametrization, but also the actual curve. Only the initial and terminal points matter. Splitting the curve into several 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.

Now, this phenomenon will be discussed in general dimensions. 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 half-space $\bar{M} = \mathbb{R}_- \times \mathbb{R}^{n-1}$, where $\mathbb{R}_- = \{(x_1, \dots, x_n) \in \mathbb{R}^n;\ x_1 \leq 0\}$. Its boundary is $\partial M = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n;\ x_1 = 0\}$. The orientation on this half-space inherited from the standard orientation is the one determined by the form $dx_2 \wedge \dots \wedge dx_n$.

The substitution $u = y/x$ then leads to

$$\begin{aligned} u'x + u &= u \ \ln u, \\ \frac{du}{dx} x &= u \ (\ln u \ - 1) \,, \\ \frac{du}{u \,(\ln u \ - 1)} &= \frac{dx}{x}, \end{aligned}$$

where $u\,(\ln u \ - 1) \neq 0$. Using another substitution, namely $t = \ln u \ - 1$, we can integrate

$$\begin{aligned} \int \frac{du}{u\,(\ln u \ - 1)} &= \int \frac{dx}{x}, \\ \int \frac{dt}{t} &= \int \frac{dx}{x}, \\ \ln|\,t\,| = \ln|\,x\,| + \ln|\,C\,|, &\quad C \neq 0, \\ \ln|\,\ln u \ - 1\,| = \ln|\,Cx\,|, &\quad C \neq 0, \\ \ln u \ - 1 = Cx, &\quad C \neq 0, \\ \ln\frac{y}{x} = Cx + 1, &\quad C \neq 0, \\ y = x\mathrm{e}^{Cx+1}, &\quad C \neq 0. \end{aligned}$$

The excluded cases $u \ = 0$ and $\ln u \ = 1$ do not lead to two more solutions since$u \ = 0$ implies $y = 0$, which cannot be put into the original equation. On the other hand, $\ln u \ = 1$ gives $y/x = \mathrm{e}$, and the function $y = \mathrm{e}x$ is clearly a solution. Therefore, the general solution is

$$y = x\mathrm{e}^{Cx+1}, \quad C \in \mathbb{R}.$$

□

8.124. Compute

$$y' = -\tfrac{4x+3y+1}{3x+2y+1}.$$

Solution. In general, we are able to solve every equation of the form

$$y' = f\left(\frac{ax + by + c}{Ax + By + C}\right). \tag{8.9}$$

If the system of linear equations

$$ax + by + c = 0, \quad Ax + By + C = 0 \tag{8.10}$$

has a unique solution x_0, y_0, then the substitution $u = x - x_0$, $v = y - y_0$ transforms the equation (||8.9||) to a homogeneous equation

$$\tfrac{dv}{du} = f\left(\tfrac{au+bv}{Au+Bv}\right).$$

If the system (||8.10||) has no solution or has infinitely many solutions, the substitution $z = ax + by$ transforms the equation (||8.9||) 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

$$\tfrac{dv}{du} = -\tfrac{4u+3v}{3u+2v},$$

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 k–dimensional manifold with a cover by compatible parametrizations φ_i. Further, let us assume that for every boundary point $x \in \partial M = \bar{M} \setminus M$, it has a neighborhood in $\bar{M}$ with parametrization $\varphi : V \subset \mathbb{R}_- \times \mathbb{R}^{k-1} \to M$ such that the points $x \in \partial M \cap \varphi(V)$ are just the image of the boundary of the half-space $\mathbb{R}_- \times \mathbb{R}^{k-1}$. The subset $\bar{M}$ with these properties is called an *oriented manifold with boundary*.

The restriction of the parametrizations including the boundary to this boundary ∂M defines a structure of a $k - 1$–dimensional oriented manifold on ∂M.

8.43. Stokes' theorem. Now, we get to a very important and useful result. The main theorem about the multidimensional analogy of curve and surface integrals will be formulated for smooth forms and smooth manifolds. A brief analysis of the proof shows that actually, we need a once continuously differentiable integrated exterior form and a twice continuously integrable 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 come into being, they are adjacent and have opposite orientations in the adjacency 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 unit decomposition subordinate to it, we can express the integrals on both sides as the sum (even a finite one, since the support of the considered from ω is compact) of integrals of forms on $\mathbb{R}^k$ or the half-space $\mathbb{R}_- \times \mathbb{R}^{k-1}$.

We can thus assume without loss of generality that $\bar{M}$ is the half-space

$$\bar{M} = \mathbb{R}_- \times \mathbb{R}^{k-1},$$

and the form ω is a form with compact support on $\bar{M}$. Then, ω will surely be the sum of the forms

$$\omega = \omega_j(x) dx_1 \wedge \cdots \wedge \hat{dx_j} \wedge \cdots \wedge dx_k,$$

where the hat indicates omission of the corresponding linear form, and $\omega_j(x)$ is a smooth function with compact support. Its exterior differential is

$$d\omega = (-1)^j \frac{\partial \omega_j}{\partial x_j} dx_1 \wedge \cdots \wedge dx_k.$$

which can be solved by further substitution $z = v/u$. We thus obtain

$$\begin{aligned} z'u\ + z &= -\frac{4+3z}{3+2z}, \\ \frac{dz}{du}u &= -\frac{2z^2+6z+4}{3+2z}, \\ \frac{2z+3}{2z^2+6z+4}\,dz &= -\frac{du}{u} \end{aligned}$$

provided $z^2 + 3z + 2 \neq 0$. Integrating, we get

$$\begin{aligned} \frac{1}{2}\ln\left|\,z^2+3z+2\,\right| &= -\ln|\,u\ \,| + \ln|\,C\,|, \quad C \neq 0, \\ \frac{1}{2}\ln\left|\,\left(z^2+3z+2\right)u^2\,\right| &= \ln|\,C\,|, \quad C \neq 0, \\ \ln\left|\,\left(z^2+3z+2\right)u^2\,\right| &= \ln C^2, \quad C \neq 0, \\ \left(z^2+3z+2\right)u^2 &= \pm C^2, \quad C \neq 0. \end{aligned}$$

We thus have

$$\left(z^2+3z+2\right)u^2 = D, \quad D \neq 0$$

and returning to the original variables,

$$\begin{aligned} \left(\frac{v^2}{u^2}+3\frac{v}{u}+2\right)u^2 &= D, \quad D \neq 0, \\ v^2+3vu+2u^2 &= D, \quad D \neq 0, \\ (y-1)^2+3(y-1)(x+1)+2(x+1)^2 &= D, \quad D \neq 0. \end{aligned}$$

Making simple rearrangements, the general solution can be expressed as

$$(x+y)\ (2x+y+1) = D, \quad D \neq 0.$$

Now, let us return to the condition $z^2 + 3z + 2 \neq 0$. It follows from $z^2 + 3z + 2 = 0$ that $z = -1$ or $z = -2$, i. e., $v = -u$ or $v = -2u$. For $v = -u$, we have $x = u\ - 1$ and $y = v\ + 1 = -u + 1$, which means that $y = -x$. Similarly, for $v = -2u$, we have $y = -2u + 1$, hence $y = -2x - 1$. However, both functions $y = -x$, $y = -2x - 1$ satisfy the original differential equations and are included in the general solution for the choice $D = 0$. Therefore, every solution is known from the implicit form

$$(x+y)\ (2x+y+1) = D, \quad D \in \mathbb{R}.$$

□

8.125. 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

$$y' = \tfrac{x^2+y^2}{2xy} = \tfrac{1+\left(\frac{y}{x}\right)^2}{2\frac{y}{x}}.$$

Using the substitution $u = y/x$, we get to the equation

$$u'x + u = \tfrac{1+u^2}{2u}.$$

If $j > 1$, the form ω on the boundary ∂M evaluates identically to zero. At the same time, invoking the fundamental theorem about antiderivatives of univariate functions, we get

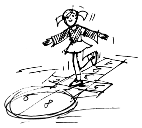
$$\begin{aligned} \int_M d\omega &= (-1)^j \int_{\mathbb{R}^{k-1}} \left(\int_{-\infty}^{\infty} \frac{\partial\omega_j}{\partial x_j}\,dx_j\right) dx_1\cdots \hat{dx_j}\cdots dx_k \\ &= (-1)^j \int_{\mathbb{R}^{k-1}} \left[\omega_j\right]_{-\infty}^{\infty} dx_1\cdots \hat{dx_j}\cdots dx_k = 0, \end{aligned}$$

since the function ω_j has compact support. So the theorem is true in this case. However, if $j = 1$, then we obtain

$$\begin{aligned} \int_M d\omega &= \int_{\mathbb{R}^{k-1}} \left(\int_{-\infty}^{0} \frac{\partial\omega_1}{\partial x_1}\,dx_1\right) dx_2\cdot\cdot\cdot\cdot\cdot\cdot dx_k \\ &= \int_{\mathbb{R}^{k-1}} \omega_1(0, x_2, \dots, x_k)dx_2\cdots dx_k = \int_{\partial M}\omega \end{aligned}$$

This finishes the proof of Stokes's theorem. □

8.44. Notes about application of Stokes' theorem. We have proved an extraordinarily strong result which covers several standard integral relations from the classic vector analysis. For instance, we can notice that by Stokes' theorem, the integration of the exterior differential $d\omega$ of any $(k-1)$–forms over a compact manifold without boundary is always zero (for example, integrating a 2–form $d\omega$ over the sphere $S^2 \subset \mathbb{R}^3$).



Let us look step by step at the cases of Stokes' theorem in lower dimensions.

The case $n = 2$**,** $k = 1$**.** We are thus examining a surface M in the plane, bounded by a curve $C = \partial M$. If we have $\omega(x, y) = f(x, y)dx + g(x, y)dy$, then $d\omega = \big(-\frac{\partial f}{\partial y} + \frac{\partial g}{\partial x}\big)dx \wedge dy$. Therefore, Stokes' theorem gives 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 so-called *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 $X = f(x, y)\frac{\partial}{\partial x} + g(x, y)\frac{\partial}{\partial y}$ defines right the form ω given above. The integral of ω_X over a curve C has the physical interpretation of the work done by movement along this curve in the force field X. Green's theorem then says, besides others, that if $\omega_X = dF$ for some function F, then the work done along a closed curve is always zero. Such fields are called potential fields and the function F is the potential of the field X.

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.

The case $n = 3, k = 2$**.** We are examining a region in $\mathbb{R}^3$, bounded by a surface S. If $\omega = f(x, y, z)dy \wedge dz + g(x, y, z)dz \wedge dx + h(x, y, z)dx \wedge dy$, we get $d\omega = \big(\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z}\big)dx \wedge dy \wedge dz$, and Stokes' theorem says that

$$\begin{aligned} &\int_S f(x,y,z)dy\wedge dz + g(x,y,z)dz\wedge dx + h(x,y,z)dx\wedge dy \\ &\quad = \int_M \left(\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z}\right)dx\wedge dy\wedge dz. \end{aligned}$$

For $u \neq \pm 1$ and $D = -1/C$, we have

$$\begin{aligned}
\frac{du}{dx}\,x &= \frac{1+u^2-2u^2}{2u},\\
\frac{2u}{1-u^2}\,du &= \frac{dx}{x},\\
-\ln\left|\,1-u^2\,\right| = \ln|\,x\,| + \ln|\,C\,|, &\quad C \neq 0,\\
\ln\frac{1}{\left|\,1-u^2\,\right|} = \ln|\,Cx\,|, &\quad C \neq 0,\\
\frac{1}{1-u^2} = Cx, &\quad C \neq 0,\\
1 = Cx\left(1-\frac{y^2}{x^2}\right), &\quad C \neq 0,\\
-\frac{D}{x} = 1-\frac{y^2}{x^2}, &\quad D \neq 0,\\
-Dx = x^2-y^2, &\quad D \neq 0.
\end{aligned}$$

The condition $u = \pm 1$ corresponds to $y = \pm x$. While $y \equiv 0$ is not a solution, both the functions $y = x$ and $y = -x$ are solutions and can be obtained by the choice $D = 0$. The general solution is thus

$$y^2 = x^2 + Dx, \quad D \in \mathbb{R}. \qquad \square$$

8.126. 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 $\mathrm{e}^{-\int a(x)\,dx}$) 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

$$\mathrm{e}^{\int \frac{2}{x^2-1}\,\mathrm{d}x} = \mathrm{e}^{\ln\left|\frac{x-1}{x+1}\right|} = \tfrac{x-1}{x+1},$$

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

$$y\tfrac{x-1}{x+1} = \int \tfrac{x(x-1)}{x+1}\,dx = \tfrac{x^2}{2} - 2x + 2\ln|\,x+1\,| + C, \quad C \in \mathbb{R}.$$

Therefore, the solutions are the functions

$$y = \tfrac{x+1}{x-1}\left(\tfrac{x^2}{2} - 2x + 2\ln|\,x+1\,| + C\right), \quad C \in \mathbb{R}.$$

As for the latter method, we first solve the corresponding homogeneous equation

$$y' = -\tfrac{2y}{x^2-1},$$

This is the statement of the so-called *Gauss–Ostrogradsky theorem*.

This theorem also has a very illustrative physical interpretation. Every vector field $X = f(x,y,z)\frac{\partial}{\partial x} + g(x,y,z)\frac{\partial}{\partial y} + h(x,y,z)\frac{\partial}{\partial z}$ defines an exterior 2–form $\omega^X(x,y,z) = f(x,y,z)dy \wedge dz + g(x,y,z)dz \wedge dx + h(x,y,z)dx \wedge dy$ by substitution for the first argument in the standard form of volume. The integral of this form over a surface can be perceived so that the integrated 2–form infinitesimally contributes, at every point to the integral, the increase equal to the volume of the parallelepiped given by the field X and a little piece of surface. If we consider the vector field to be the velocity of movement of the particular points of the space, this will be the "flow rate" through the given surface. On the right-hand side of the integral, there is an expression which can be defined as $d(\omega^X) = (\operatorname{div} X)dx \wedge dy \wedge dz$. Gauss–Ostrogradsky theorem says that if $\operatorname{div} X$ equals zero identically, then the total flow rate through the boundary surface of the region is zero as well. Such fields, with $\operatorname{div} X = 0$, are called solenoidal vector fields.

The case $n = 3$, $k = 1$. In this case, we have a surface M in $\mathbb{R}^3$ bounded by a curve C. If the linear form ω is the differential of some function, we find out that the integral over the surface depends on the boundary curve only. This is the *classical Stokes' theorem*. If we use the standard scalar product, just like in the plane, to identify the vector field $X = f\frac{\partial}{\partial x} + g\frac{\partial}{\partial y} + h\frac{\partial}{\partial z}$ with the form $\omega = f dx + g dx + h dz$, we obtain

$$\int_C f dx + g dx + h dz = \int_M d\omega,$$

where $d\omega = \left(\frac{\partial h}{\partial y} - \frac{\partial g}{\partial z}\right)dy \wedge dz + \left(\frac{\partial f}{\partial z} - \frac{\partial g}{\partial x}\right)dz \wedge dx + \left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y}\right)dx \wedge dy$. This 2–form can again be identified with a single vector field $\operatorname{rot} X$, which yields $d\omega$ by substitution into the standard form of volume. This field is called the *rotation or curl of the vector field X*. We can see that in the three-dimensional space, vector fields X having the property that $\omega_X = dF$ for some function F are given by the condition $\operatorname{rot} X = 0$. They are called conservative (or potential) vector fields.

3. Differential equations

In this section, we will get back to (vector) functions of one variable, which will be given and examined in terms of their instantaneous changes. At the end, we will stop for a while to look at equations containing partial derivatives.

8.45. 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, we once defined differences for the same reason, and it was just the relations between the values of the quantities and the changes of them or other quantities which lead to the so-called difference equations. As a motivating introduction to equations containing derivatives of unknown functions, we will now return to the difference equations for a while.

The simplest model was interests of deposits or loans (and the same for the so-called Malthusian model of populations). The increase was proportional to the value, see 1.10. Considering continuous modeling, the same request leads to an equation connecting which is an equation with separated variables. We have

$$\begin{aligned} \frac{dy}{dx} &= -\frac{2y}{x^2-1}, \\ \frac{dy}{y} &= -\frac{2}{x^2-1}\,dx, \\ \ln|\,y\,| = -\ln|\,x-1\,| + \ln|\,x+1\,| &+ \ln|\,C\,|, \quad C \neq 0, \\ \ln|\,y\,| &= \ln\left|C\,\frac{x+1}{x-1}\right|, \quad C \neq 0, \\ y &= C\,\frac{x+1}{x-1}, \quad C \neq 0, \end{aligned}$$

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

$$y = \tfrac{C(x+1)}{x-1}, \quad C \in \mathbb{R}.$$

Now, we will consider the constant C to be a function $C(x)$. Differentiating leads to

$$y' = \tfrac{C'(x)\,(x+1)(x-1)+C(x)\,(x-1)-C(x)\,(x+1)}{(x-1)^2}.$$

Substituting this into the original equation, we get

$$\tfrac{C'(x)\,(x+1)(x-1)+C(x)\,(x-1)-C(x)\,(x+1)}{(x-1)^2} = x - \tfrac{2\,C(x)\,(x+1)}{(x-1)(x^2-1)}.$$

It follows that

$$C'(x) = \tfrac{x(x-1)}{x+1},$$

i. e.,

$$\begin{aligned} C(x) &= \int \frac{x(x-1)}{x+1}\,dx, \\ C(x) &= \frac{x^2}{2} - 2x + 2\ln|\,x+1\,| + C, \quad C \in \mathbb{R}. \end{aligned}$$

Now, it suffices to substitute:

$$y = C(x)\,\tfrac{x+1}{x-1} = \tfrac{x+1}{x-1}\left(\tfrac{x^2}{2} - 2x + 2\ln|\,x+1\,| + C\right), \quad C \in \mathbb{R}.$$

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.

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

$$y = C\mathrm{e}^{\int a(x)\,dx}, \quad C \in \mathbb{R}.$$

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)

$$y = \mathrm{e}^{\int_{x_0}^{x} a(t)\,dt}\left(y_0 + \textstyle\int_{x_0}^{x} b(t)\,\mathrm{e}^{-\int_{x_0}^{t} a(s)\,ds}\,\mathrm{d}t\right).$$

Let us remark that the linear equation has no singular solution, and the general solution contains a $C \in \mathbb{R}$. □

8.127. Solve the linear equation

$$\left(y' + 2xy\right)\mathrm{e}^{x^2} = \cos x.$$

the derivative $y'(t)$ of a function with its value

$$y'(t) = r \cdot y(t) \tag{8.4}$$

with a proportionality constant r.

It is easy to guess the *solution* of this equation, i. e. a function $y(t)$ which satisfies the equality identically,

$$y(t) = C\,\mathrm{e}^{rt}$$

with an arbitrary constant C. This constant can be determined uniquely be choosing the so-called *initial values* $y_0 = y(t_0)$ at some point t_0. If a part of the increase in our model were given by a constant action independent of the value y or t (like bank charges or the natural decrease of population as a result of sending some part of it to slaughterhouses), we could use an equation with a constant s on the right-hand side.

$$y'(t) = r \cdot y(t) + s. \tag{8.5}$$

Apparently, the solution of this equation is the function

$$y(t) = C\,\mathrm{e}^{rt} - \frac{s}{r}.$$

It is very easy to come across this solution if we realize that the set of all solutions of the equation (8.4) is a one-dimensional vector space, while the solutions of the equation (8.5) are obtained by adding any one of its solutions to the solutions of the previous equation. We can then easily find the constant solution $y(t) = k$ for $k = -\frac{s}{r}$.

Similarly, in paragraph 1.13, we managed to create the so-called logistic model of population growth based upon 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. We also wanted the model to behave similarly 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

$$p'(t) = p(t)\left(-\frac{r}{K}p(t) + r\right), \tag{8.6}$$

i. e., at the value $p(t) = K$ for a large constant K, the instantaneous increase of the function p is indeed zero, while for $p(t)$ near zero, the ratio of the rate of increase of the population and its size is close to r, which is often a small number (roughly hundredths) expressing the rate of increase of the population in good conditions.

It is surely not easy to solve such an equation without knowing the proper theory (although we will be able to deal with this type of equations presently). However, as an exercise on differentiation, we can easily verify that the following function is a solution for every constant C:

$$p(t) = \frac{K}{1 + CK\,\mathrm{e}^{-rt}}.$$

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 $y\,\mathrm{e}^{x^2}$. Thus, we can immediately calculate

$$\begin{aligned}\left(y\,\mathrm{e}^{x^2}\right)' &= \cos x,\\ y\,\mathrm{e}^{x^2} &= \int \cos x\,dx,\\ y\,\mathrm{e}^{x^2} &= \sin x + C, \quad C \in \mathbb{R},\\ y &= \mathrm{e}^{-x^2}\left(\sin x + C\right), \quad C \in \mathbb{R}.\end{aligned}$$

□

8.128. Find all non-zero solutions of the Bernoulli equation

$$y' - \tfrac{y}{x} = 3xy^2 .$$

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

$$u' + \tfrac{u}{x} = -3x.$$

Similarly to the previous exercise, we have

$$u = \mathrm{e}^{-\ln|x|}\left[\textstyle\int -3x\,\mathrm{e}^{\ln|x|}\,dx\right] ,$$

where $\ln|x|$ was obtained as an (arbitrary) antiderivative to $1/x$. Furhter,

$$\begin{aligned}u &= \mathrm{e}^{\ln\frac{1}{|x|}}\left[\int -3x\,\mathrm{e}^{\ln|x|}\,dx\right] ,\\ u &= \frac{1}{|x|}\left[\int -3x\,|x|\,dx\right] .\end{aligned}$$

The absolute value can be replaced with a sign that can be canceled, i. e., it suffices to consider

$$u = \tfrac{1}{x}\left[\textstyle\int -3x^2\,dx\right] = \tfrac{1}{x}\left[-x^3 + C\right] , \quad C \in \mathbb{R}.$$

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). □

8.129. Interchanging the variables, solve the equation

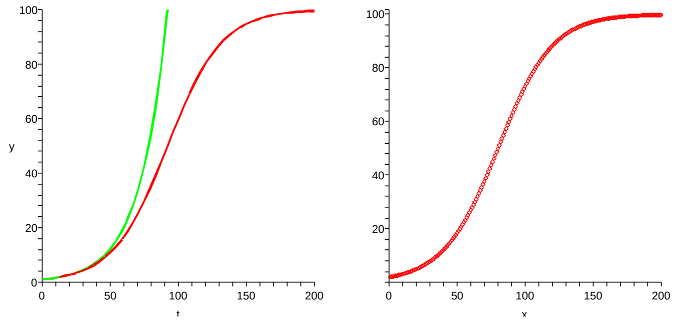
$$y\,dx - \left(x + y^2 \sin y\right)\,dy = 0.$$

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:



Confronting the red graph (left-hand picture) of this function with the choice $K = 100$, $r = 0,05$, and $C = 1$ (the first two were used in 1.13 this way, the last one roughly corresponds to the initial value $p(0) = 1$) with the right-hand picture (the solution of the difference equation from 1.13 with the same values of the parameters), we can see that both approaches to population modeling indeed yield quite similar results. To compare the output, the left-hand picture also contains in green the graph of the solution of the equation (8.4) with the same constant r and initial condition.

8.46. First-order differential equations. By an (ordinary) first-order differential equation, we usually mean 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

$$F(y'(t), y(t), t) = 0.$$

The writing resembles implicitly given functions $y(t)$; however, this time, there is a dependency upon the derivative of the wanted function $y(t)$.

If the equation is solved at least explicitly with regard to the derivative, i. e.,

$$y'(t) = f(t, y(t))$$

for some function $f : \mathbb{R}^2 \to \mathbb{R}$, we can imagine graphically what this equation defines. For every value (t, y) in the plane, we can consider the arrow corresponding to the vector $(1, f(t, y))$, i. e., the velocity with which the point of the graph of the solution moves through the plane in dependence on the free parameter t.

For the equation (8.6), for instance, we get the following picture (illustrating the solution for the initial condition as above).

$$\frac{dy}{dx} = \frac{y}{x + y^2 \sin y},$$
$$\frac{dx}{dy} = \left(\frac{y}{x + y^2 \sin y}\right)^{-1} = \frac{x}{y} + y \sin y,$$
$$x' = \frac{1}{y} x + y \sin y.$$

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 **??**.

L. Practical problems leading to differential equations

8.130. A water purification plant with volume 2000 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 m^3/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 dt, $\frac{m}{V} \cdot v\,\mathrm{d}t$ 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

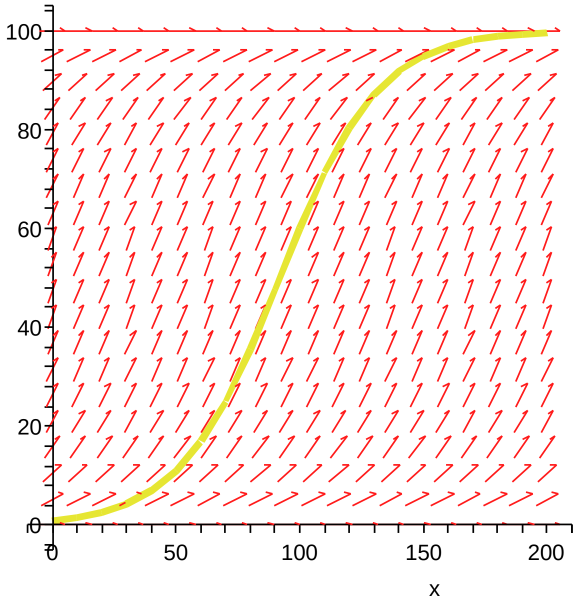
$$\frac{\mathrm{d}m}{m} = -\frac{v}{V}\,\mathrm{d}t.$$

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 $t \doteq 6\,\mathrm{h}\ 35\,\mathrm{min}$. □

8.131. 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,

$$\frac{df}{dt} = k(P - f(t)).$$



Considering these pictures, we can intuitively anticipate that for every initial condition, there will exist a unique solution of our equation. However, as we will see, this proposition holds only for sufficiently smooth functions f.

8.47. Integration of differential equations. Before examining existence of the solutions of the differential equations, we present at least one truly elementary method of solution. It transforms the solution to ordinary integration, which usually leads to an implicit description of the solution.

Equations with separated variables

Consider a differential equation in the form

$$y'(t) = f(t) \cdot g(y(t)) \tag{8.7}$$

for two continuous functions of a real variable, f and g.

The solution of this equation can be obtained by integration, i. e., we find the antiderivatives

$$G(y) = \int \frac{dy}{g(y)}, \quad F(x) = \int f(x)dx.$$

This procedure reliably finds a solution which satisfies $g(y(t)) \neq 0$.

Then, computing the function $y(x)$ from the implicitly given formula $F(x) + C = G(y)$ with an arbitrary constant C leads to the solution, because differentiating this equation using the chain rule for the composite function $G(y(x))$ indeed leads to $\frac{1}{g(y)} \cdot y'(x) = f(x)$.

As an example, we can find the solution of the equation

$$y'(x) = x \cdot y(x).$$

Direct calculation gives $\ln|y(x)| = \frac{1}{2}x^2 + C$. Hence it looks (at least for positive values of y) as

$$y(x) = \mathrm{e}^{\frac{1}{2}x^2 + C} = D \cdot \mathrm{e}^{\frac{1}{2}x^2},$$

where D is an arbitrary positive constant now. Let us stop for a while to examine the resulting formula and signs thoroughly. The

Separating the variables and introducing a constant K (the number of people who know the message at time $t = 0$ must be $P - K$), we get the solution

$$f(t) = P - Ke^{-kt},$$

where k is a positive real constant.

Apparently, this model makes sense for large values of P only. □

8.132. The speed at which an epidemic spreads in a given closed population consisting of P people is directly proportional to the product of the number of people who have been infected and the number of people who have not. Determine the function $f(t)$ describing the number of infected people in time.

Solution. Just like in the previous problem, we construct a differential equation:

$$\frac{df}{dt} = k \cdot f(t)\,(P - f(t))\,.$$

Again, separating the variables and introducing suitable constants K and L, we obtain

$$f(t) = \frac{K}{1 + Le^{-Kkt}}.$$

□

8.133. 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. □

8.134. The acceleration of an object falling in a constant gravitational field with a certain resistance of the environment is given by the formula

$$\frac{dv}{dt} = g - kv,$$

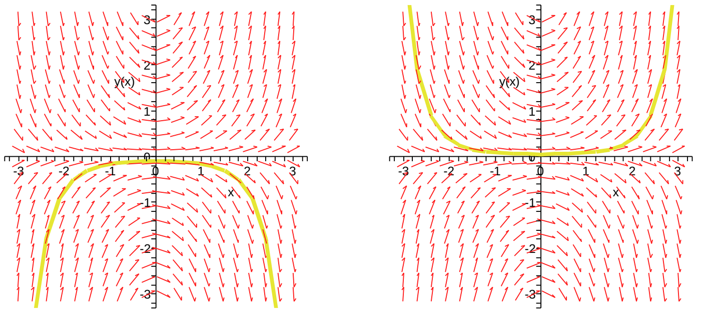
where k is a constant which expresses the resistance of the environment. An object was dropped in a gravitational field with $g = 10$ ms^{-2} at the initial speed of 5 ms^{-1}, the resistance constant is $k = 0.5$ s^{-1}. What will the speed of the object be in three seconds?

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. □

constant solution $y(x) = 0$ satisfies our equation as well, and for negative values of y, we can use the same solution with negative constants D. In fact, the constant D can be arbitrary, and we have found a solution satisfying any initial value.



The picture shows two solutions which demonstrate the instability of the equation with regard to the initial values: If, for any x_0, we change a tiny y_0 from a negative value to a positive one, then the behavior of the resulting solution changes dramatically. Moreover, we should notice the constant solution $y(x) = 0$, which satisfies the initial condition $y(x_0) = 0$.

Using separation of variables, we can easily solve the non-linear equation from the previous paragraph which described a logistic population model. Try this as an exercise.

In the first chapter, we paid much attention to the so-called linear difference equations, and their general solution, looking quite awful, was determined in paragraph 1.10 on page 13. Although it was clear beforehand that it will be a one-dimensional affine space of satisfying sequences, it was a hardly transparent sum, because we needed to take into account all of the changing coefficients.

We can thus use this as a source of inspiration for the following construction of the solution of a general first-order linear equation

$$y'(t) = a(t)y(t) + b(t) \tag{8.8}$$

with continuous coefficients $a(t)$ ans $b(t)$.

First of all, let us find the solution of the homogenized equation $y'(t) = a(t)y(t)$. This can be computed easily by separation of variables, obtaining

$$y(t) = y_0 F(t, t_0), \quad F(t, s) = \mathrm{e}^{\int_s^t a(x)\,dx}\,.$$

In the case of difference equations, we "guessed" the solution, and then we proved by induction that it was correct. It is even simpler now, as it suffices to differentiate the correct solution to verify the statement.

The solution of first-order linear equations

The solution of the equation (8.8) with initial values $y(t_0) = y_0$ is (locally in a neighborhood 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 $F(t, s) = \mathrm{e}^{\int_s^t a(x)\,dx}$.

Verify the correctness of the solution by yourselves (pay proper attention to the differentiation of the integral where t is both in the upper bound and a free parameter in the integrand).

8.135. The rate of increase of a population of a certain type of bug is indirectly proportional to its size. At time $t = 0$, the population had 100 bugs. In a month, the population doubled. What will the size of the population be in two months?

Solution. Let us consider a continuous approximation of the number of bugs, and let their amount be denoted by P. Then, we can build the following equation:

$$\frac{dP}{dt} = \frac{k}{P},$$

$P = \sqrt{Kt + c}$. Substituting the given values, we get $P(2) = \sqrt{7}\cdot 100$, which is an estimate of the actual number of bugs. □

8.136. Find the equation of the curve with the following properties: It lies in the first quadrant, goes through the point $[1, 3/4]$, and its tangent at any point marks on the positive half-axis y a segment whose length is the same as the distance of that point from the origin. ○

8.137. Consider a chemical compound C isolated in a container. C is unstable, with half-time of a molecule equal to q time units. If there were M moles of the compound C in the container at the beginning (i. e., at time $t = 0$), how many moles of it will be there at time $t \geq 0$? ○

8.138. A 100-gram body lengthens a spring of 5 cm if hung on it. Express the dependency of its position on time t provided the speed of the body is 10 cm/s when going through the equilibrium point. ○

Further practical problems that lead to differential equations can be found on page **??**.

M. Higher-order differential equations

8.139. 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

$$x'(t) = -y(t), \quad y'(t) = x(t)$$

from 8.7. 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.

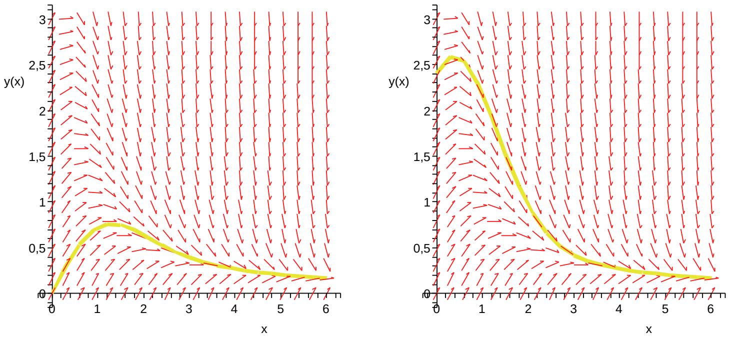
Therefore, in order to determine a unique solution, we need to know not only the initial position y_0, but also the speed of the motion at that moment. These two pieces of information uniquely determine both the amplitude and the initial phase.

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

Now, we can, for instance, directly solve the equation

$$y'(x) = 1 - x \cdot y(x),$$

this time encountering stable behavior, visible in the following picture.



8.48. Transformation of coordinates. Our pictures tend to indicate that differential equations can be perceived as geometric objects (the "directional field of the arrows"), so we should be able to look for the solution by conveniently chosen coordinates. We will get back to this point of view later; now, we will only show three simple and typical tricks as they seem from the explicit form of the equations in coordinates.

We begin with the so-called *homogeneous equations* of the form

$$y'(t) = f\big(\frac{y(t)}{t}\big).$$

Considering a transformation $z = \frac{y}{t}$, assuming that $t \neq 0$, then we get by the chain rule that

$$z'(t) = \frac{1}{t^2}\big(t\,y'(t) - y(t)\big) = \frac{1}{t}(f(z) - z),$$

which is an equation with separated variables.

Another example is the so-called *Bernoulli differential equations*, which are of the form

$$y'(t) = f(t)y(t) + g(t)y(t)^n,$$

where $n \neq 0, 1$. The choice of the transformation $z = y^{1-n}$ leads to the equation

$$\begin{aligned} z'(t) &= (1-n)y(t)^{-n}(f(t)y(t) + g(t)y^n) \\ &= (1-n)f(t)z(t) + (1-n)g(t), \end{aligned}$$

which is a linear equation, which we are able to integrate.

In the end, let us take a look at an extraordinarily important equation, the so-called *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 we are able to guess a particular solution $x(t)$. Then, we can use the transformation

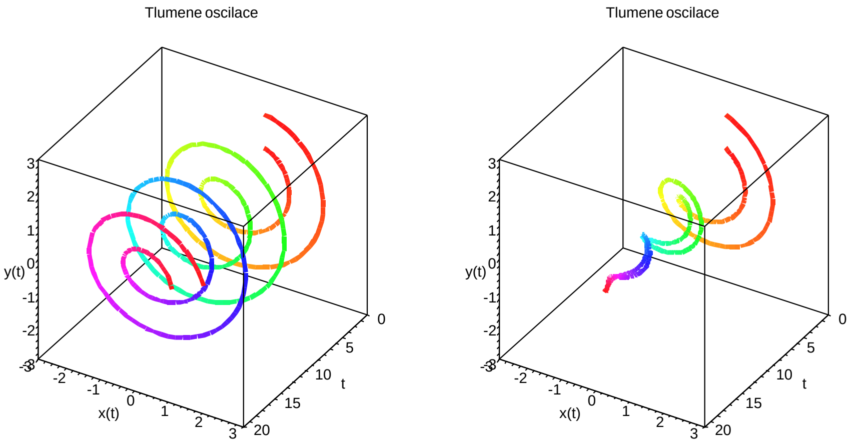
$$z(t) = \frac{1}{y(t) - x(t)}.$$

Verify by yourselves that this transformation leads to the equation

$$z'(t) = -\big(f(t) + 2x(t)g(t)\big)z(t) - g(t).$$

$$y''(t) = -y(t) - \alpha y'(t),$$

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.

8.140. 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 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

$$y(t) = -\tfrac{1}{2}\sin(2t) + \tfrac{1}{4}t\sin(2t) + \mathcal{L}^{-1}\left(e^{-\pi s}\frac{s}{(s^2+4)^2}\right).$$

Just as we saw in the case of integration of functions (which is, in fact, the simplest type of equations with separated variables), the equations usually do not have a solution expressible explicitly in terms of elementary functions.

Similarly as with standard engineer tables of values of special functions, books listing the solutions of basic equations were compiled as well.[4] Today, the wisdom concealed in them is essentially transferred to software systems like Maple or Mathematica. There, we can assign any task on ordinary differential equations, and we will get the results in a surprisingly good deal of cases, yet after all, it will not be possible for most problems.

The way out of this is numerical methods, which try only to approximate the solutions. However, to be able to use them, we still need good theoretical starting points regarding existence, uniqueness, and stability of the solutions.

We begin with the so-called Picard–Lindelöf theorem:

Existence and uniqueness of the solutions of ODEs

8.49. Theorem. *Let a function $f(t, y) : \mathbb{R}^2 \to \mathbb{R}$ have continuous partial derivatives on an open set U. Then for every point $(t_0, y_0) \in U \supset \mathbb{R}^2$, there exists a maximal interval $I = [t_0 - a, t_0 + b]$, with positive $a, b \in \mathbb{R}$, and a unique function $y(t) : I \to \mathbb{R}$ which is a solution of the equation*

$$y'(t) = f(t, y(t))$$

on the interval I.

Proof. Notice that if a function $y(t)$ is a solution of our equation satisfying the initial condition $y(t_0) = t_0$, then it also satisfies the equation

$$y(t) = y_0 + \int_{t_0}^{t} y'(t)\,dt = y_0 + \int_{t_0}^{t} f(t, y(t))\,dt.$$

However, the right-hand side of this expression is, up to constant, the integral operator

$$L(y)(t) = y_0 + \int_{t_0}^{t} f(t, y(t))\,dt.$$

When solving our first-order differential equations, we are thus looking for a fixed point of this operator L, i. e., we want 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 $L(y)$, then it immediately follows from the antiderivative theorem that $y(t)$ indeed satisfies the given differential equation, including the initial conditions.

We can quite easily guess for the operator L how much its values $L(y)$ and $L(z)$ differ for various arguments $y(t)$ and $z(t)$. Indeed, thanks to the partial derivatives of the function f being continuous, we know that f is locally Lipschitz. This means that we have the bound

$$|f(t, y) - f(t, z)| \le C|y - z|,$$

with a constant C if we restrict the values (t, y) to a neighborhood of the point (t_0, y_0) with compact closure. We choose an $\varepsilon > 0$ and restrict the value of t to some interval $J = [t_0 - a_0, t_0 + b_0]$ so

[4] E. g., Kamke ...

However, by formula (‖7.36‖), 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

$$y(t)=\begin{cases}-\frac{1}{2}\sin(2t)+\frac{1}{4}t\sin(2t) & \text{for } 0\le t<\pi\\ \frac{\pi-2}{4}\sin(2t) & \text{for } t\ge\pi\end{cases}$$

□

8.141. 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

$$y^{(n)}+a_1y^{(n-1)}+a_2y^{(n-2)}+\cdots+a_{n-1}y'+a_ny=f(x)$$

for certain constants $a_1,\dots,a_n\in\mathbb{R}$. Moreover, we have $f(x)\equiv 0$, i. e., the equation is homogeneous.

First of all, we will find the roots of the so-called characteristic polynomial

$$\lambda^n+a_1\lambda^{n-1}+a_2\lambda^{n-2}+\cdots+a_{n-1}\lambda+a_n.$$

Each real root λ with multiplicity k corresponds to the k solutions

$$e^{\lambda x},\, x\,e^{\lambda x},\dots,x^{k-1}\,e^{\lambda x}$$

and every pair of complex roots $\lambda=\alpha\pm i\beta$ with multiplicity k corresponds to the k pairs of solutions

$$\begin{gathered}e^{\alpha x}\cos(\beta x)\ ,\,x\,e^{\alpha x}\cos(\beta x)\ ,\dots,x^{k-1}\,e^{\alpha x}\cos(\beta x)\ ,\\ e^{\alpha x}\sin(\beta x)\ ,\,x\,e^{\alpha x}\sin(\beta x)\ ,\dots,x^{k-1}\,e^{\alpha x}\sin(\beta x)\ .\end{gathered}$$

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:

$$y=C_1e^{4x}+C_2x\,e^{4x}+C_3e^{-3x},\qquad C_1,C_2,C_3\in\mathbb{R}.$$

□

8.142. Compute

$$y'''+y''+9y'+9y=e^x+10\cos(3x)\ .$$

Solution. First, we will solve the corresponding homogeneous equation. The characteristic polynomial is equal to

$$\lambda^3+\lambda^2+9\lambda+9,$$

with roots $\lambda_1=-1$, $\lambda_2=3i$, $\lambda_3=-3i$. The general solution of the corresponding homogeneous equation is thus

$$y=C_1e^{-x}+C_2\cos(3x)+C_3\sin(3x)\ ,\quad C_1,C_2,C_3\in\mathbb{R}.$$

The solution of the non-homogeneous equation is of the form

$$y=C_1e^{-x}+C_2\cos(3x)+C_3\sin(3x)+y_p,\quad C_1,C_2,C_3\in\mathbb{R}$$

for a particular solution y_p of the non-homogeneous equation.

The right-hand side of the given equation is of a special form. In general, if the non-homogeneous part is given by a function

that $J\times[y_0-\varepsilon,\,y_0+\varepsilon]\subset U$, and we consider only those functions $y(t)$ and $z(t)$ which, for $t\in J$, satisfy

$$\max_{t\in J}|y(t)-y_0|<\varepsilon,\ \max_{t\in J}|z(t)-y_0|<\varepsilon.$$

Now,we obtain the bound

$$\begin{aligned}|(L(y)-L(z))(t)| &= \left|\int_{t_0}^{t}f(t,y(t))-f(t,z(t))\,dt\right|\\ &\le\int_{t_0}^{t}|f(t,y(t))-f(t,z(t))|\,dt\\ &\le C\int_{t_0}^{t}|y(t)-z(t)|\,dt\\ &\le D|t-t_0|\end{aligned}$$

for suitable constants C and D. For a sufficiently small $\delta>0$, we thus have

$$\max_{|t-t_0|<\delta}|L(y)(t)-L(z)(t)|<\max_{|t-t_0|<\delta}c\,|y(t)-z(t)|$$

for some constant $0<c<1$. In paragraph 7.19 on page 441, these operators were called *contraction*. However, for the assumptions of the Banach contraction theorem, which guarantees a uniquely determined fixed point, we need even completeness of the space X of functions on which the operator L works.

In our case, we can notice that merely from the continuity of the mapping $f(t,y)$, there follows a uniform bound for all of the functions $y(t)$ considered above and the values $t>s$ in their domain:

$$|L(y)(t)-L(y)(s)|\le\int_s^t|f(t,y(t)|dt\le d\,|t-s|$$

with a universal constant $d>0$. Therefore, besides the conditions mentioned above, we can even restrict ourselves to the subset of all uniformly continuous functions. This set is already compact, hence it is a complete set of continuous functions on our interval, see Arzela-Askoli theorem 7.23. Therefore, there exists a unique fixed point $y(t)$ of this contraction L, which is the solution of our equation.

It remains to show the existence of a maximal interval $I=(t_0-a,t_0+b)$. Let us suppose that we have found a solution $y(t)$ on an interval (t_0,t_1), and, at the same time, the limit

$$y_1=\lim_{t\to t_1}y(t)$$

exists and is finite. Then, it follows from what has been proved above that there must exist a solution with initial condition (t_1,y_1), in some neighborhood of the point t_1, and on the left-hand side of it, it must coincide with the solution $y(t)$. Therefore, the solution $y(t)$ can surely be extended on the right-hand side of t_1. There are thus only two possibilities when the solution right of t_1 does not exist: either there is no finite limit $y(t)$ at t_1 from the left, or the limit y_1 exists, yet the point (t_1,y_1) is on the boundary of the domain U of the function f. In both cases, we indeed have a maximal extension of the solution right of t_0.

The argumentation for the maximal solution left of t_0 is analogous. □

$$P_n(x)\, \mathrm{e}^{\alpha x},$$

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

$$\mathrm{e}^{\alpha x}\, [P_m(x) \cos(\beta x) + S_n(x) \sin(\beta x)] \ ,$$

where P_m is a polynomial of degree m and S_n is a polynomial of degree n, there exists a particular solution of the form

$$y_p = x^k\, \mathrm{e}^{\alpha x}\, [R_l(x) \cos(\beta x) + T_l(x) \sin(\beta x)] \ ,$$

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

$$y_{p_2}(x) = x\,[B\cos(3x) + C\sin(3x)]\,.$$

Recall that the number $\lambda = 3\mathrm{i}$ was obtained as a root of the characteristic polynomial. We can easily compute the derivatives

$$\begin{aligned} y'_{p_2}(x) &= [B\cos(3x) + C\sin(3x)] \\ &\quad + x\,[-3B\sin(3x) + 3C\cos(3x)]\,, \\ y''_{p_2}(x) &= 2\,[-3B\sin(3x) + 3C\cos(3x)] \\ &\quad + x\,[-9B\cos(3x) - 9C\sin(3x)]\,, \\ y'''_{p_2}(x) &= 3\,[-9B\cos(3x) - 9C\sin(3x)] \\ &\quad + x\,[27B\sin(3x) - 27C\cos(3x)]\,. \end{aligned}$$

Substituting them into the equation, whose right-hand side contains the function $y = 10\cos(3x)$, we get

$$-18B\cos(3x) - 18C\sin(3x) - 6B\sin(3x) + 6C\cos(3x) = 10\cos(3x)\ .$$

Confronting the coefficients leads to the system of linear equations

$$-18B + 6C = 10, \quad -18C - 6B = 0$$

with the only solution $B = -1/2$ and $C = 1/6$, i. e.,

$$y_{p_2}(x) = x\left[-\tfrac{1}{2}\cos(3x) + \tfrac{1}{6}\sin(3x)\right].$$

Altogether, the general solution is

$$y = C_1\mathrm{e}^{-x} + C_2\cos(3x) + C_3\sin(3x) + \frac{1}{20}\mathrm{e}^x$$
$$-\frac{1}{2}x\cos(3x) + \frac{1}{6}x\sin(3x)\ , \quad C_1, C_2, C_3 \in \mathbb{R}.$$

8.50. 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. By a concrete bound for the constant c from the proof, we can get even straight bounds for the errors. Try to think this out as an exercise (see the proof of Banach fixed-point theorem in paragraph 7.19). It can then be shown quite easily and directly that it is a uniformly convergent sequence of continuous functions, so the limit will again be a continuous function (without invoking the complicated theorems from the seventh chapter).

Picard's approximations

The unique solution of the equation

$$y'(t) = f(t, y(t))$$

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 (the so-called *Picard's approximation*):

$$y_0(t) = y_0, \quad y_{n+1}(t) = L(y_n),\ n = 1, \dots.$$

It is a uniformly converging sequence of continuous functions with continuous limit $y(t)$.

Let us notice that we actually needed only the Lipschitzness of partial derivatives of the function, so the theorem holds with this weaker assumption as well. We will show in the next paragraph that mere continuity of the function f guarantees the existence of the solution as well, yet it is insufficient for the uniqueness.

8.51. Ambiguity of solutions. Let us begin with a really simple example. Consider the equation

$$y'(t) = \sqrt{|y(t)|}\,.$$

Separating the variables, we can easily find the solution

$$y(t) = \frac{1}{4}(t + C)^2,$$

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 will also be locally exactly one solution. The solution must apparently keep being non-decreasing, hence for negative values y_0, we get the same solution, only with the other sign and $t + C < 0$.

However, for the initial condition $(t_0, y_0) = (t_0, 0)$, we have 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 bound arbitrarily, see the picture. 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 continuous solution of the equation*

$$y'(t) = f(t, y(t))$$

locally in some neighborhood of t_0.

Proof. The proof will be presented only roughly, leaving the

details to the reader. Instead of using Picard's approximations, we will proceed quite naively.

□

8.143. 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

$$c_1e^{-x} + c_2e^{-2x},$$

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 $c_1e^{-x} + c_2e^{-2x} - xe^{-2x}$, $c_1, c_2 \in \mathbb{R}$. □

8.144. Determine the general solution of the equation

$$y'' + y' = 1.$$

Solution. The characteristic polynomial of the given equation is $x^2 + x$, with roots 0 and -1. Therefore, the general solution of the homogenized equation is $c_1 + c_2e^{-x}$, where $c_1, c_2 \in \mathbb{R}$.

We are looking for a particular solution in the form ax, $a \in \mathbb{R}$ (since zero is a root of the characteristic polynomial). Substituting into the original equation, we get $a = 1$. The general solution of the given non-homogeneous equation is $c_1 + c_2e^{-x} + x$, $c_1, c_2 \in \mathbb{R}$. □

8.145. Determine the general solution of the equation

$$y'' + 5y' + 6y = e^{-2x}.$$

Solution. The characteristic polynomial of the equation is $x^2 + 5x + 6 = (x + 2)(x + 3)$, its roots are -2 and -3. The general solution of the homogenized equation is thus $c_1e^{-2x} + c_2e^{-3x}$, $c_1, c_2 \in \mathbb{R}$. We are looking for a particular solution in the form axe^{-2x}, (-2 is a root of the characteristic polynomial), $a \in \mathbb{R}$, using the method of undetermined coefficients. Substitution into the original equation yields $a = 1$. Hence, the general solution of the given equation is

$$c_1e^{-2x} + c_2e^{-3x} + xe^{-2x}.$$

We will construct a solution to the right of the initial point t_0. To this purpose, we select a small step $h > 0$ and label the points

$$t_k = t_0 + kh, \quad k = 1, 2, \dots.$$

The value of the derivative $f(t_0, y_0)$ of the corresponding curve of the solution $(t, y(t))$ is defined at the initial point (t_0, y_0), so we can substitute a parametrized line with the same derivative:

$$y^{(0)}(t) = y_0 + f(t_0, y_0)(t - t_0),$$

and we label $y_1 = y^{(0)}(t_1)$. We thus inductively construct 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, we define $y_h(t)$ by gluing the particular linear parts, i. e.,

$$y_h(t) = y^{(k)}(t) \quad \text{for all } t \in [kh, (k + 1)h].$$

This is clearly a continuous function, which is called *Euler's approximation* of the solution.

Now it "only" remains to prove that the limit of the functions y_h for h approaching zero exists and is a solution.

For this, we need to notice (as we have already done in the proof of the theorem on uniqueness and existence of the solution) that, thanks to $f(t, y)$ being uniformly continuous on the neighborhood U where we are looking for a solution, we have, for any selected $\varepsilon > 0$, a δ such that

$$|f(t, y) - f(s, z)| < \varepsilon,$$

whenever $\|(t - s, y - z)\| < \delta$.

Especially, all of out functions y_h are in the set of uniformly continuous functions on the concerned interval. Therefore, by Arzela-Askoli theorem (see paragraph 7.23 on page 443), there exists a sequence of values $h_n \to 0$ such that the corresponding sequence of functions y_{h_n} converges uniformly to a continuous function $y(n)$. Further, let us write more simply $y_n(t) = y_{h_n} \to y(t)$.

However, for each of the continuous functions y_h, we have only finitely many points in the interval $[t_0, t]$ where it is not differentiable, so we can write

$$y_n(t) = y_0 + \int_{t_0}^{t} y_n'(s)\,ds.$$

On the other hand, the derivatives on the particular intervals are constant, so we can write (here, k is the largest such that $t_0 + kh_n \le t$, while y_j and t_j are the points from the definition of the function y_{h_n})

$$y_n(t) = y_0 + \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(t_j, y_j)ds$$
$$+ \int_{t_k}^{t} f(t_k, y_k).$$

Instead, we would like to see

$$y_n(t) = y_0 + \int_{t_0}^{t} f(s, y_n(s))\,ds,$$

but the difference between this integral and the last two terms in the previous expression is bound by the possible differences of the function $f(t, y)$ and the lengths of the intervals. Thanks to our

□

8.146. 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

$$c_1 + c_2 e^x - 5x.$$

□

8.147. Solve the equation

$$y'' - 2y' + y = \tfrac{\mathrm{e}^x}{x^2+1}.$$

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

$$\begin{aligned}
C_1'(x)\, y_1(x) + \cdots + C_n'(x)\, y_n(x) &= 0,\\
C_1'(x)\, y_1'(x) + \cdots + C_n'(x)\, y_n'(x) &= 0,\\
&\ \vdots\\
C_1'(x)\, y_1^{(n-2)}(x) + \cdots + C_n'(x)\, y_n^{(n-2)}(x) &= 0,\\
C_1'(x)\, y_1^{(n-1)}(x) + \cdots + C_n'(x)\, y_n^{(n-1)}(x) &= f(x).
\end{aligned}$$

The roots of the characteristic polynomial $\lambda^2 - 2\lambda + 1$ are $\lambda_1 = \lambda_2 = 1$. Therefore, we are looking for the solution in the form

$$C_1(x)\,\mathrm{e}^x + C_2(x)\, x\,\mathrm{e}^x,$$

considering the system

$$\begin{aligned}
\breve{A}C_1'(x)\,\mathrm{e}^x + C_2'(x)\, x\,\mathrm{e}^x &= 0,\\
C_1'(x)\,\mathrm{e}^x + C_2'(x)\left[\mathrm{e}^x + x\,\mathrm{e}^x\right] &= \frac{\mathrm{e}^x}{x^2+1}.
\end{aligned}$$

We can compute the unknowns $C_1'(x)$ and $C_2'(x)$ using Cramer's rule. It follows from

$$\begin{vmatrix} \mathrm{e}^x & x\,\mathrm{e}^x \\ \mathrm{e}^x & \mathrm{e}^x + x\,\mathrm{e}^x \end{vmatrix} = \mathrm{e}^{2x},$$

$$\begin{vmatrix} 0 & x\,\mathrm{e}^x \\ \frac{\mathrm{e}^x}{x^2+1} & \mathrm{e}^x + x\,\mathrm{e}^x \end{vmatrix} = -x\,\frac{\mathrm{e}^{2x}}{x^2+1},$$

$$\begin{vmatrix} \mathrm{e}^x & 0 \\ \mathrm{e}^x & \frac{\mathrm{e}^x}{x^2+1} \end{vmatrix} = \frac{\mathrm{e}^{2x}}{x^2+1}$$

universal bound for $f(t, y)$ above, we can thus use just the last integral instead of the actual values in the limit process $\lim_{n\to\infty} y_n(t)$, thereby obtaining

$$\begin{aligned}
y(t) &= \lim_{n\to\infty}\Big(y_0 + \int_{t_0}^{t} f(s, y_n(s))\, ds\Big)\\
&= y_0 + \int_{t_0}^{t} \big(\lim f(s, y_n(s))\big)\, ds\\
&= y_0 + \int_{t_0}^{t} f(s, y(s))\, ds,
\end{aligned}$$

where we used the uniform convergence $y_n(t) \to y(t)$.

This proves the theorem. □

8.52. Systems of first-order equations. The problem of finding the solution of the equation $y'(x) = f(x, y)$ can also be viewed as looking for a (parametrized) curve $(x(t), y(t))$ in the plane where we have fixed the parametrization of the variable $x(t) = t$ beforehand. However, if we accept this point of view, then we can forget this fixed choice for one variable and we can add an arbitrary number of variables.

In the plane, for instance, we can write such a system in the form

$$x'(t) = f(t, x(t), y(t)), \qquad y'(t) = g(t, x(t), y(t))$$

with two functions $f, g : \mathbb{R}^3 \to \mathbb{R}$. Similarly for more variables.

A simple example in the plane might be the system of equations

$$x'(t) = -y(t), \qquad y'(t) = x(t).$$

It can be easily guessed (or verified at least) that there is a solution of this system,

$$x(t) = R\cos t, \qquad y(t) = R\sin t,$$

with an arbitrary non-negative constant R, and the curves of the solution will be exactly the parametrized circles with radius R.

In the general case, we will work with the vector notation of the system in the form

$$x'(t) = f(t, x(t))$$

for a vector function $x : \mathbb{R} \to \mathbb{R}^n$ and a mapping $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$. We are able to extend the validity of the theorem on uniqueness and existence of the solution to such systems:

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 $(t_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}$, 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*

$$\begin{aligned}
x_1'(x) &= f_1(t, x_1(t), \ldots, x_n(x))\\
&\ \vdots\\
x_n'(x) &= f_n(t, x_1(t), \ldots, x_n(x))
\end{aligned}$$

with initial condition

$$x_1(t_0) = x_1, \ldots, x_n(t_0) = x_n.$$

that

$$C_1(x) = -\int \frac{x}{x^2+1}\,dx = -\frac{1}{2}\ln\left(x^2+1\right) + C_1, \qquad C_1 \in \mathbb{R},$$

$$C_2(x) = \int \frac{dx}{x^2+1} = \arctan x + C_2, \qquad C_2 \in \mathbb{R}.$$

Hence, the general solution is

$$y = C_1 \mathrm{e}^x + C_2 x\,\mathrm{e}^x - \tfrac{1}{2}\,\mathrm{e}^x \ln\left(x^2+1\right) + x\,\mathrm{e}^x \arctan x, \qquad C_1, C_2 \in \mathbb{R}.$$

□

8.148. 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,

$$\frac{2}{9}e^{2x} + \frac{5}{18}e^{-x} + \frac{1}{3}xe^{-x} - \frac{1}{2}e^x.$$

□

Further problems concerning higher-order differential equations can be found on page **??**

N. Applications of the Laplace transform

Differential equations with constant coefficients can also be solved using the Laplace transform.

8.149. Let $\mathcal{L}(y)(s)$ denote the Laplace transform of a function $y(t)$. Integrating by parts, prove that

Solution.

$$\mathcal{L}(y')(s) = s\mathcal{L}(y)(s) - y(0) \tag{8.11}$$

$$\mathcal{L}(y'')(s) = s^2\mathcal{L}(y) - sy(0) - y'(0)$$

and, by induction:

$$\mathcal{L}(y^{(n)})(s) = s^n\,\mathcal{L}(y)(s) - \textstyle\sum_{i=1}^{n} s^{n-i}\,y^{(i-1)}(0)\,.$$

□

8.150. 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 ||8.149|| that

$$s^2\,\mathcal{L}(y)(s) + 4\mathcal{L}(y)(s) = \mathcal{L}(\sin 2t)(s).$$

We also have

$$\mathcal{L}(\sin 2t)(s) = \frac{2}{s^2+4},$$

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 we showed in Theorem 8.49. The unknown function $x(t) = (x_1(t), \dots, x_n(t))$ is a curve in $\mathbb{R}^n$ satisfying the given equation, so its components $x_i(t)$ are again expressible in terms of integrals

$$x_i(t) = x_i(t_0) + \int_{t_0}^{t} x_i'(t)\,dt = x_i + \int_{t_0}^{t} f_i(t, x(t))\,dt.$$

We are thus working with the integral operator $y \mapsto L(y)$ once again, this time mapping curves in $\mathbb{R}^n$ to curves in $\mathbb{R}^n$, and we are looking for its fixed point. Since the Euclidean distance of two points in $\mathbb{R}^n$ is always bounded from above by the sum of the sizes of the differences of the particular components, the proof goes on in much the same way as in the case 8.49. We only need to notice that the size of the vector

$$\|f(t, z_1, \dots, z_n) - f(t, y_1, \dots, y_n)\|$$

is bounded from above by the sum

$$\begin{aligned}&\|f(t, z_1, \dots, z_n) - f(t, y_1, z_2 \dots, z_n)\| + \dots\\ &\qquad + \|f(t, y_1, \dots, y_{n-1}, z_n) - f(t, y_1, \dots, y_n)\|.\end{aligned}$$

We recommend to go through the proof of Theorem 8.49 from this point of view and to think out the details. □

When we introduced and examined models of a real system, the so-called qualitative behavior of the solution in dependence on the initial conditions and free parameters of the system (i. e. constants or functions) is essential.

As a quite simple example of a system of first-order equations, we can notice the standard population model "predator – prey", which 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. We assume that the increase 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, we assume that the predator would only naturally die out (i. e. exponential decrease with coefficient γ). Further, we 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. We get a system of two equations:

Lotka-Volterra model

$$\begin{aligned} x'(t) &= \alpha x(t) - \beta y(t)x(t)\\ y'(t) &= -\gamma y(t) + \delta\beta x(t)y(t).\end{aligned}$$

It is interesting that the same model captures quite well the progress of unemployment in the system limited to employers and their employees, considering the employees to be the predators, while the employers play the role of the prey.

Vlozit cca ctyri obrazky znazornujici dynamiku pro ruzne hodnoty koeficientu a opatrit komentarem!!!!!

Much information about this and other models can be found in literature.

i. e.,

$$\mathcal{L}(y)(s) = \frac{2}{(s^2+4)^2}.$$

The inverse transform leads to

$$y(t) = \tfrac{1}{8}\sin 2t - \tfrac{1}{4}t\cos 2t\ . \qquad \square$$

8.151. 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

$$50 = (As+B)(s+3)^2 + C(s^2+1)(s+3) + D(s^2+1).$$

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

$$\begin{aligned}\mathcal{L}(y)(s) &= \tfrac{-3s+4}{s^2+1} + \tfrac{3}{s+3} + \tfrac{5}{(s+3)^2} + \tfrac{1}{s+3} + \tfrac{7}{(s+3)^2}\\ &= \tfrac{-3s}{s^2+1} + \tfrac{4}{s^2+1} + \tfrac{4}{s+3} + \tfrac{12}{(s+3)^2}.\end{aligned}$$

Now, the inverse Laplace transform yields the solution in the form

$$y(t) = -3\cos t + 4\sin t + 4e^{-3t} + 12te^{-3t}\ . \qquad \square$$

8.53. Stability of systems of equations. Now, we restrict ourselves to a single basic theorem about stability of systems. We can notice that assuming that the partial derivatives of the functions defining the systems are continuous (in fact, Lipschitz) guarantees the continuity of the solution's behavior in dependence on the initial conditions as well as the equations themselves.

However, as the distance of t from the initial value t_0 increases, the bounds grow exponentially! Therefore, this result has only a local usage, and it is in no contradiction with the example of the unstably behaving equation $y'(t) = t\ y(t)$ illustrated in paragraph 8.47.[5]

Let us consider two systems of equations written in the vector form

$$x'(t) = f(t, x(t)), \quad y'(t) = g(t, y(t))$$

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 we can label the finite values

$$C = \sup_{x\neq y;\ (t,x),\ (t,y)\in U} \frac{|f(t,x) - f(t,y)|}{|x-y|}$$

$$B = \sup_{(t,x)\in U} |f(t,x) - g(t,x)|$$

Having this notation, we can formulate our fundamental theorem:

Theorem. *Let $x(t)$ and $y(t)$ be two fixed solutions*

$$x'(t) = f(t, x(t)), \quad y'(t) = g(t, y(t))$$

of the systems considered above, given by initial conditions $x(t_0) = x_0$ and $y(t_0) = y_0$. Then,

$$|x(t) - y(t)| \le |x_0 - y_0|\, \mathrm{e}^{C|t-t_0|} + \tfrac{B}{C}\big(\mathrm{e}^{C|t-t_0|} - 1\big).$$

Proof. Without loss of generality, we can assume that $t_0 = 0$. From the expression of the solutions $x(t)$ and $y(t)$ as fixed points of the corresponding integral operators, we immediately get the bound

$$|x(t) - y(t)| \le |x_0 - y_0| + \int_0^t |f(s, x(s)) - g(s, y(s))|\, ds.$$

The integrand can be further bound as follows:

$$\begin{aligned}&|f(s, x(s)) - g(s, y(s))|\\ &\quad\le |f(s, x(s)) - f(s, y(s))| + |f(s, y(s)) - g(s, y(s))\\ &\quad\le C\,|x(s) - y(s)| + B\end{aligned}$$

If we denote $F(t) = |x(t) - y(t)|$, $\alpha = |x_0 - y_0|$, we can write our bound as

$$F(t) \le \alpha + \int_0^t (C\,F(s) + B)\, ds.$$

Such a bound can be quite easily used thanks to the following general result, which is known as *Gronwall's inequality*. Notice the similarity with the general solution of linear equations.

Lemma. *Let a real-valued function $F(t)$ satisfy, for the input values from an interval $t \in [0, t_{max}]$,*

$$F(t) \le \alpha(t) + \int_0^t \beta(s)F(s)\, ds$$

[5] Much more information can be found in a nice book Teschl ...

8.152. 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

$$\mathcal{L}\left(f''\right)(s) = s^2\mathcal{L}(f)(s) - s\lim_{t\to 0+} f(t) - \lim_{t\to 0+} f'(t)$$

and

$$\mathcal{L}(\cos(bt))(s) = \tfrac{s}{s^2+b^2}, \quad b \in \mathbb{R}.$$

Applying the Laplace transform to the given differential equation then gives

$$s^2\mathcal{L}(y)(s) - sc_1 - c_2 = \tfrac{s}{s^2+\pi^2} - \mathcal{L}(y)(s),$$

i. e.,

$$\mathcal{L}(y)(s) = \frac{s}{\left(s^2+1\right)\left(s^2+\pi^2\right)} + \frac{c_1 s}{s^2+1} + \frac{c_2}{s^2+1}. \tag{8.12}$$

Therefore, it suffices to find a function y which satisfies (‖8.12‖). Performing partial fraction decomposition, we obtain

$$\tfrac{s}{\left(s^2+1\right)\left(s^2+\pi^2\right)} = \tfrac{1}{\pi^2-1}\left(\tfrac{s}{s^2+1} - \tfrac{s}{s^2+\pi^2}\right).$$

The above expression of $\mathcal{L}(\cos(bt))(s)$ and the already proved formula

$$\mathcal{L}(\sin t)(s) = \tfrac{1}{s^2+1}$$

then yield the wanted solution

$$y(t) = \tfrac{1}{\pi^2-1}(\cos t - \cos(\pi t)) + c_1\cos t + c_2\sin t\,. \qquad \square$$

8.153. 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

$$\mathcal{L}\left(\mathrm{e}^{\pm t}\right)(s) = \tfrac{1}{s\mp 1},$$

transforms the first equation to

$$\begin{aligned} s^2\mathcal{L}(x)(s) - s\lim_{t\to 0+} x(t) - \lim_{t\to 0+} x'(t) + s\mathcal{L}(x)(s) - \lim_{t\to 0+} x(t) &= \\ = \mathcal{L}(y)(s) - \left(s^2\mathcal{L}(y)(s) - s\lim_{t\to 0+} y(t) - \lim_{t\to 0+} y'(t)\right) &+ \tfrac{1}{s-1} \end{aligned}$$

and the second one to

$$\begin{aligned} s\mathcal{L}(x)(s) - \lim_{t\to 0+} x(t) + 2\mathcal{L}(x)(s) &= \\ = -\mathcal{L}(y)(s) + s\mathcal{L}(y)(s) - \lim_{t\to 0+} y(t) &+ \tfrac{1}{s+1}. \end{aligned}$$

Evaluating the limits (according to the initial conditions), we obtain the linear equations

$$s^2\mathcal{L}(x)(s) - 1 + s\mathcal{L}(x)(s) = \mathcal{L}(y)(s) - s^2\mathcal{L}(y)(s) + \tfrac{1}{s-1}$$

and

$$s\mathcal{L}(x)(s) + 2\mathcal{L}(x)(s) = -\mathcal{L}(y)(s) + s\mathcal{L}(y)(s) + \tfrac{1}{s+1}$$

with the only solution

$$\mathcal{L}(x)(s) = \tfrac{2s-1}{2(s-1)(s+1)^2}, \quad \mathcal{L}(y)(s) = \tfrac{3s}{2\left(s^2-1\right)^2}.$$

Once again, we perform partial fraction decomposition, getting

$$\mathcal{L}(x)(s) = \tfrac{1}{8}\tfrac{1}{s-1} + \tfrac{3}{4}\tfrac{1}{(s+1)^2} - \tfrac{1}{8}\tfrac{1}{s+1} = \tfrac{3}{4}\tfrac{1}{(s+1)^2} + \tfrac{1}{4}\tfrac{1}{s^2-1}.$$

for some real-valued functions $\alpha(t)$, $\beta(t)$, where $\beta(t) \ge 0$. Then, we also have

$$F(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s)\,\mathrm{e}^{\int_s^t \beta(r)\,dr}\;ds$$

for all $t \in [0, t_{max}]$. If we even have that $\alpha(t)$ is non-decreasing, then

$$F(t) \le \alpha(t)\,\mathrm{e}^{\int_0^t \beta(s)\,ds}\;.$$

Proof of the lemma. For the sake of transparency, let us write

$$G(t) = \mathrm{e}^{-\int_0^t \beta(s)\,ds}\;.$$

Using the first assumption of the theorem, direct computation yields

$$\begin{aligned} \frac{d}{dt}\left(G(t)\int_0^t \beta(s)F(s)\,ds\right) &= \\ &= \beta(t)G(t)\left(F(t) - \int_0^t \beta(s)F(s)\,ds\right) \\ &\le \alpha(t)\beta(t)G(t) \end{aligned}$$

Now, 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, we can continue:

$$F(t) \le \alpha(t)\big(1 + \int_0^t \beta(s)\,\mathrm{e}^{\int_s^t \beta(r)\,dr}\;ds\big).$$

Now, it suffices to notice that the integrand is actually a derivative:

$$-\beta(s)\,\mathrm{e}^{\int_s^t \beta(r)\,dr} = \frac{d}{ds}\left(\mathrm{e}^{\int_s^t \beta(r)\,dr}\right),$$

so we finally obtain

$$\begin{aligned} F(t) &\le \alpha(t)\big(1 - \int_0^t \frac{d}{ds}\,\mathrm{e}^{\int_s^t \beta(r)\,dr}\;ds\big) \\ &= \alpha(t)\big(1 + \mathrm{e}^{\int_s^t \beta(r)\,dr} - 1\big), \end{aligned}$$

and the second proposition of the lemma has thus been proved as well. $\square$

Now, we can finish the proof of the theorem about continuous dependency upon the parameters. We have already obtained the bound $F(t) \le \alpha + \int_0^t (C\,F(s) + B)\,ds$, and using a merely modified function $\tilde{F}(t) = F(t) + \frac{B}{C}$, it yields

$$\tilde{F}(t) \le \tfrac{D}{C} + \alpha + \int_0^t C\tilde{F}(s)\,ds.$$

This is the assumption of Gronwall's inequality with (even) constant parameters, so by the second proposition of the lemma, we get

$$F(t) + \tfrac{B}{C} \le (\alpha + \tfrac{B}{C})\,\mathrm{e}^{\int_0^t C\,ds},,$$

which is the statement

$$F(t) \le \alpha\,\mathrm{e}^{Ct} + \tfrac{B}{C}(\mathrm{e}^{Ct} - 1)$$

we have wanted to prove. $\square$

Since we have already computed that

$$\mathcal{L}\left(t\,\mathrm{e}^{-t}\right)(s) = \tfrac{1}{(s+1)^2}, \quad \mathcal{L}\left(\sinh t\right)(s) = \tfrac{1}{s^2-1},$$
$$\mathcal{L}\left(t\sinh t\right)(s) = \tfrac{2s}{(s^2-1)^2},$$

we get

$$x(t) = \tfrac{3}{4}\,t\,\mathrm{e}^{-t} + \tfrac{1}{4}\sinh t, \qquad y(t) = \tfrac{3}{4}\,t\sinh t.$$

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 $s > 1$). □

8.154. Find the solution of the following system of differential equations:

$$\begin{aligned} x'(t) &= -2x(t) + 3y(t) + 3t^2, \\ y'(t) &= -4x(t) + 5y(t) + e^t, \quad x(0) = 1,\ y(0) = -1 \end{aligned}$$

Solution.

$$\begin{aligned} \mathcal{L}(x')(s) &= \mathcal{L}(-2x + 3y + 3t^2)(s), \\ \mathcal{L}(y')(s) &= \mathcal{L}(-4x + 5y + e^t)(s). \end{aligned}$$

The left-hand sides can be written using (||8.11||), while the right-hand sides can be rewritten thanks to linearity of the $\mathcal{L}$ operator. Since $\mathcal{L}(3t^2)(s) = \frac{6}{s^3}$ and $\mathcal{L}(e^t)(s) = \frac{1}{s-1}$, we get the system of linear equations

$$\begin{aligned} s\mathcal{L}(x)(s) - 1 &= -2\mathcal{L}(x)(s) + 3\mathcal{L}(y)(s) + \tfrac{6}{s^3}, \\ s\mathcal{L}(y)(s) + 1 &= -4\mathcal{L}(x)(s) + 5\mathcal{L}(y)(s) + \tfrac{1}{s-1}. \end{aligned}$$

In matrices, this is $\mathbf{A}(s)\hat{\mathbf{x}}(s) = \mathbf{b}(s)$, where

$$\mathbf{A}(s) = \begin{pmatrix} s+2 & -3 \\ 4 & s-5 \end{pmatrix}, \hat{\mathbf{x}}(s) = \begin{pmatrix} \mathcal{L}(x)(s) \\ \mathcal{L}(y)(s) \end{pmatrix} \text{ and } \mathbf{b}(s) = \begin{pmatrix} 1 + \frac{6}{s^3} \\ -1 + \frac{1}{s-1} \end{pmatrix}.$$

Cramer's rule says that

$$\mathcal{L}(x)(s) = \tfrac{|\mathbf{A}_1|}{|\mathbf{A}|}, \quad \mathcal{L}(y)(s) = \tfrac{|\mathbf{A}_2|}{|\mathbf{A}|}, \text{ where}$$

$$|\mathbf{A}| = \begin{vmatrix} s+2 & -3 \\ 4 & s-5 \end{vmatrix} = s^2 - 3s + 2,$$

$$|\mathbf{A}_1| = \begin{vmatrix} 1 + \frac{6}{s^3} & -3 \\ -1 + \frac{1}{s-1} & s-5 \end{vmatrix} = (s-5)(1 + \tfrac{6}{s^3}) + 3(-1 + \tfrac{1}{s-1})$$

$$|\mathbf{A}_2| = \begin{vmatrix} s+2 & 1 + \frac{6}{s^3} \\ 4 & -1 + \frac{1}{s-1} \end{vmatrix} = (s+2)(-1 + \tfrac{1}{s-1}) - 4 - \tfrac{24}{s^3}.$$

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:

$$\mathcal{L}(x)(s) = -\tfrac{39}{2s^2} - \tfrac{3}{(s-1)^2} + \tfrac{28}{s-1} - \tfrac{21}{4(s-2)} - \tfrac{15}{s^3} - \tfrac{87}{4s},$$

$$\mathcal{L}(x)(s) = -\tfrac{18}{s^2} - \tfrac{3}{(s-1)^2} + \tfrac{27}{s-1} - \tfrac{7}{s-2} - \tfrac{12}{s^3} - \tfrac{21}{s}.$$

The continuous dependency upon both the initial conditions and the potential further parameters in which the function f would be Lipschitz-continuous immediately follows from the statement of the theorem. A really simple equation in one variable $x'(t) = x(t)$ with exponential solution shows that we cannot hope in better general results.

8.54. Differentiability of the solutions. In practical problems, we are often interested in the differentiability of the obtained solutions, especially with regard to the initial conditions or other parameters of the system.

We can notice that in the general vector notation of the system of ordinary equations

$$x'(t) = f(t, x(t)),$$

we can always suppose that the vector function does not depend implicitly on t. Indeed, if it depends explicitly on t, we can add another variable x_0 and write the same system of equations for the curve $\tilde{x}'(t) = (x_0(t), x_1(t), \dots, x_n(t))$ as

$$\begin{aligned} x_0'(t) &= 1 \\ x_1'(t) &= f_1(x_0(t), x_1(t), \dots, x_n(t)) \\ &\vdots \\ x_n'(t) &= f_n(x_0(t), x_1(t), \dots, x_n(t)) \end{aligned}$$

with initial conditions

$$x_0(t_0) = t_0,\ x(t_0) = x_1, \dots, x_n(t_0) = x_n.$$

Such systems, which do not explicitly depend on time, are called *autonomous systems of ordinary differential equations*.

To simplify the method, we will deal with autonomous systems dependent on parameters λ and with initial conditions

$$y'(t) = f(y(t), \lambda),\ y(t_0) = x. \tag{8.9}$$

Without loss of generality, we will, in the case of autonomous systems, always consider the initial value $t_0 = 0$, and should need arise, we will 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), the solution will always be once more differentiable than the function f. This can be easily derived inductively by applying the chain rule. If f is continuously differentiable,

$$y''(t) = D^1 f(y(t)) \cdot y'(t) = D^1 f(y(t)) \cdot f(y(t))$$

exists and is continuous. Having all the derivatives up to order two continuous, we get the expression for the third derivative:

$$\begin{aligned} y^{(3)}(t) &= D^2 f(y(t))\big(f(y(t)), f(y(t))\big) \\ &\quad + \big(D^1 f(y(t))\big)^2 \cdot f(y(t)). \end{aligned}$$

Think out the argumentation for higher orders in detail.

Let us assume for a while that there is a solution $y(t, x)$ of our system (8.9) which is continuously differentiable in the parameters $x \in \mathbb{R}^n$ as well. Then, the derivative

$$\Phi(t, x) = D_x^1(y(t, x)),$$

i. e. the Jacobian matrix of all partial derivatives with respect to the coordinates x_i, which depends on the time t as well as the initial

Now, the inverse transform yields the solution of this Cauchy problem:

$$x(t) = -\tfrac{39}{2}t - 3te^t + 28e^t - \tfrac{21}{4}e^{2t} - \tfrac{15}{2}t^2 - \tfrac{87}{4},$$
$$y(t) = -18t - 3te^t + 27e^t - 7e^{2t} - 6t^2 - 21\,. \qquad \square$$

O. Equation of heat conduction

8.155. Find the solution to the so-called equation of heat conduction (equation of diffusion)

$$u_t(x,t) = a^2\, u_{xx}(x,t), \quad x \in \mathbb{R},\ t > 0$$

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

$$\mathcal{F}\,(u_t)\,(\omega,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} u_t(x,t)\,\mathrm{e}^{-i\omega x}\,dx =$$
$$= \left(\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} u\,(x,t)\,\mathrm{e}^{-i\omega x}\,dx \right)',$$

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

$$\mathcal{F}\left(a^2\, u_{xx}\right)(\omega,t) = a^2\,\mathcal{F}\,(u_{xx})\,(\omega,t) = -a^2\omega^2\,\mathcal{F}\,(u)\,(\omega,t).$$

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

$$y(\omega,t) = K\,(\omega)\,\mathrm{e}^{-a^2\omega^2 t}, \quad K\,(\omega) \in \mathbb{R}.$$

It remains to determine $K(\omega)$. The transformation of the initial condition gives

$$\mathcal{F}\,(f)\,(\omega) = \lim_{t\to 0+} \mathcal{F}\,(u)\,(\omega,t) = \lim_{t\to 0+} y(\omega,t) = K\,(\omega)\,\mathrm{e}^0 = K\,(\omega),$$

hence

$$y(\omega,t) = \mathcal{F}\,(f)\,(\omega)\,\mathrm{e}^{-a^2\omega^2 t}, \quad K\,(\omega) \in \mathbb{R}.$$

Now, using the inverse Fourier transform, we can return to the original differential equation with solution

$$u\,(x,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} y(\omega,t)\,\mathrm{e}^{i\omega x}\,d\omega =$$
$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathcal{F}\,(f)\,(\omega)\,\mathrm{e}^{-a^2\omega^2 t}\,\mathrm{e}^{i\omega x}\,d\omega =$$

condition x, can be determined using the chain rule:

$$\begin{aligned} D_x^1\big(y'(t,x)\big) &= \frac{d}{dt}\big(D_x^1 y(t,x)\big) \\ &= D^1 f(y(t,x)) \cdot D_x^1 y(t,x). \end{aligned}$$

The derivatives with respect to the initial conditions along the solution $y(t,x)$ of the system (8.54) are thus given as the solutions of a system of n^2 first-order equations with initial condition

$$\text{(8.10)} \qquad \Phi'(t,x) = F(t,x) \cdot \Phi(t,x),\ \ \Phi(0,x) = E,$$

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.

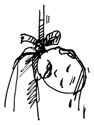
The following theorem says that for systems of continuously differentiable right-hand sides f, the derivatives with respect to the parameters can indeed be obtained in this way.

Differentiability of the solutions

Theorem. *Let us 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 upon a parameter $\lambda \in \mathbb{R}^k$ with initial condition at a point $x \in U$*

$$y'(t) = f(y(t),\lambda), \quad y(0) = x$$

has a unique solution $y(t,x,\lambda)$, which is a mapping with continuous first derivatives with respect to each variable.



Proof. First, we can notice that we can consider a system dependent on parameters to be an ordinary autonomous system with no parameters if we consider even the parameters to be space variables and we add (vector) conditions $\lambda'(t) = 0$ and $\lambda(0) = \lambda$. Therefore, without loss of generality, we can prove the theorem for autonomous systems with no further parameters and concentrate on the dependency upon the initial conditions.

Just like in the case of the fundamental existence theorem, we will build upon Picard's approximations of the solution using the integral operator

$$y_0(t,x) = x,\ \ y_{k+1}(t,x) = x + \int_0^t f(y_k(s,x))\,ds.$$

Merely specifying the proof of this theorem 8.49, we can verify the uniform convergence of the approximations $y_k(t,x)$ to the solution $y(t,x)$, including the variable x.

Now, for the initial condition, let us fix a point x_0 and a small neighborhood V of its, which, should need be, will be reduced during the following bounds, and let us write C for the constant which, thanks to Lipschitzness of the function f, gives the bound

$$|f(y) - f(z)| \le C\,|y - z|$$

on this neighborhood. We have already known that if the derivative

$$\Phi(t,x) = D_x^1 y(t,x)$$

of the solution $y(t,x)$ exists, then it will be given by the equation (8.10) with an initial condition. Therefore, let us define $\Phi(t,x)$ by this equation and examine the expression

$$G(t,h) = |y(t,x_0+h) - y(t,x_0) - h\Phi(t,x_0)|$$

$$= \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s)\, \mathrm{e}^{-i\omega s}\, ds \right) \ \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{i\omega x}\, d\omega =$$

$$= \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} f(s) \left(\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{-i\omega(s\ -x)}\, d\omega \right)\ ds.$$

Computing the Fourier transform $\mathcal{F}(f)$ of the function $f(t) = \mathrm{e}^{-at^2}$ for $a > 0$, we have obtained (while relabeling the variables)

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-cp^2}\, \mathrm{e}^{-irp}\, dp = \tfrac{1}{\sqrt{2c}}\, \mathrm{e}^{-\frac{r^2}{4c}}, \quad c > 0.$$

According to this formula (consider $c = a^2 t > 0$, $p = \omega$, $r = s - x$), we have

$$\tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \mathrm{e}^{-a^2\omega^2 t}\, \mathrm{e}^{-i\omega(s\ -x)}\, d\omega = \tfrac{1}{\sqrt{2a^2 t}}\, \mathrm{e}^{-\frac{(s\ -x)^2}{4a^2 t}},$$

Therefore,

$$u\ (x, t) = \tfrac{1}{2a\sqrt{\pi t}} \int\limits_{-\infty}^{\infty} f(s)\, \mathrm{e}^{-\frac{(x-s)^2}{4a^2 t}}\, ds.$$

□

P. Numerical solution of differential equations

Now, we present two simple exercises on applying the Euler method for solving differential equations.

8.156. 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.
$h_2 = 0.25$ $y^{(2)} = (1.0000 \quad 0.7500 \quad 0.6094 \quad 0.5165 \quad 0.4498 \quad 0.3992 \quad 0.3594 \quad 0.3271 \quad 0.3004)$ Maximal difference: 0.1094.
$h_3 = 0.125$ $y^{(3)} = (1.0000 \quad 0.8750 \quad 0.7793 \quad 0.7034 \quad 0.6415 \quad 0.5901 \quad 0.5466 \quad 0.5092 \quad 0.4768 \quad 0.4484 \quad 0.4233 \quad 0.4009 \quad 0.3808 \quad 0.3627 \quad 0.3462 \quad 0.3312 \quad 0.3175)$ Maximal difference: 0.0322.

with small increases $h \in \mathbb{R}^n$. In order to prove that the continuous derivative exists, we just have to show that

$$\lim_{h \to 0} \frac{1}{h}\, G(t, h) = 0.$$

We will need several bounds to this purpose. First, from the last theorem about continuous dependence upon initial conditions, we can immediately see the bound

$$|y(t, x_0 + h) - y(t, x_0)| \le |h|\ \mathrm{e}^{C|t|}\ .$$

In the next step, we use Taylor's expansion with remainder for the mapping f,

$$f(y) - f(z) = D^1 f(z) \cdot (y - z) + R(y, z),$$

where $R(y, z)$ satisfies $|R(y, z)|/|y - z| \to 0$ for $|y - z| \to 0$. We get the first bound which uses the definition of the mapping $\Phi(t, x_0)$ using its derivative, and we write $F(t, x) = D^1 f(y(t, x))$ again.

$$\begin{aligned} G(t, h) \le \int_0^t & |f(y(s, x_0 + h)) - f(y(s, x_0)) \\ & - h\ F(s, x_0)\Phi(s, x_0)|\, ds \\ \le \int_0^t & \|F(s, x_0)\|\ |y(s, x_0{+}h) - y(s, x_0) - h\ \Phi(s, x_0)|\, ds \\ + \int_0^t & |R(y(s, x_0 + h), y(s, x_0))|\, ds, \end{aligned}$$

where we are working with the norm on matrices given as the maximum of the absolute values of their entries.

Since we assume that $F(t, x)$ is continuous, we can bound the norm in our neighborhood V and for $|t| < T$ with a sufficiently small T to remain in the neighborhood V by

$$\|F(t, x_0)\| \le B,$$

and, at the same time, for any selected constant $\varepsilon > 0$, we can find a bound $|h| < \delta$ for which the remainde R satisfies

$$\begin{aligned} |R(y(t, x_0 + h), y(t, x_0))| \ & \le \varepsilon |y(t, x_0 + h) - y(t, x_0)| \\ & \le |h| \varepsilon\, \mathrm{e}^{CT}\ . \end{aligned}$$

Therefore, our bound can be improved as follows:

$$G(t, h) \le B \int_0^t G(s, h)\, ds + \varepsilon |h| T\, \mathrm{e}^{CT}\ .$$

Gronwall's lemma now gives

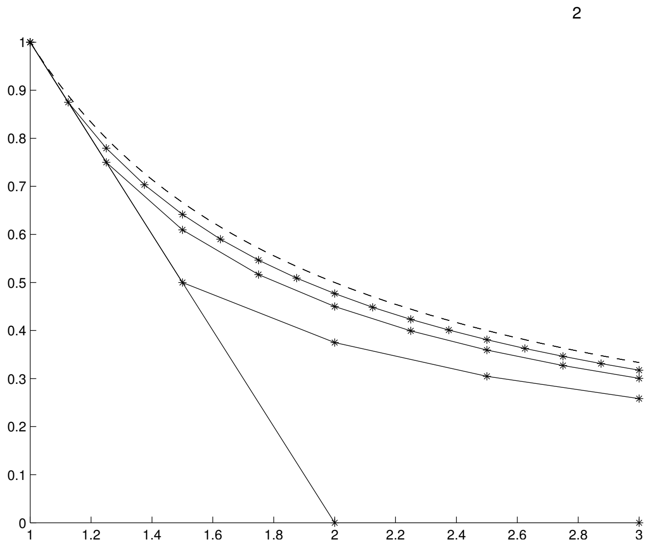
$$G(t, h) \le \varepsilon |h| T\, \mathrm{e}^{(C+B)T}\ .$$

However, hence it follows that $\lim_{h \to 0} \frac{1}{h} G(t, h)$ converges to zero, which finishes the proof. □

It can be proved very similarly that continuous differentiability of the right-hand side up to order k (inclusive) guarantees the same order of differentiability of the solution in all input parameters.

It even holds that if the right-hand side f is analytic in all parameters, then the dependency of the solution on the parameters is analytic as well.

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 $y = 1/x$.



□

8.157. 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.

8.55. Flows of vector fields. Before going to higher-order equations, we stop for a while to look at systems first-order of equations from the geometrical point of view. Now, we can geometrically formalize the right-hand side of an autonomous system as assigning the vector $f(x) \in \mathbb{R}^n$ in the direction space of the Euclidean space $\mathbb{R}^n$ to each of its points x in the considered domain. We talk about the *vector field* $X(x) = f(x)$.

If a vector field X on an open set $U \subset \mathbb{R}^n$ is given, then we can define for every differentiable function f on U its derivative in the direction of the vector field X by

$$X(f) : U \to \mathbb{R}, \qquad X(f)(x) = d_{X(x)}f.$$

Therefore, if we have, in coordinates, $X(x) = (X_1(x), \dots X_n(x))$, then

$$X(f)(x) = X_1(x)\frac{\partial f}{\partial x_1}(x) + \dots + X_n(x)\frac{\partial f}{\partial x_n}(x).$$

The simplest vector fields will have all coordinate functions equal to zero except for one function X_i which will be constantly equal to one. Such a field then corresponds to the partial derivative with respect to the variable x_i. This is also matched by the common notation

$$X(x) = X_1(x)\frac{\partial}{\partial x_1} + \dots + X_n(x)\frac{\partial}{\partial x_n}.$$

Now, the problem of finding the solution of our system of equations can be described equivalently as looking for a curve which satisfies

$$x'(t) = X(x(t))$$

for each t in its domain. The tangent vector of the wanted 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

$$\mathrm{Fl}^X_t : \mathbb{R}^n \to \mathbb{R}^n,$$

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 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 further directly guarantees that

$$\mathrm{Fl}^X_{t+s}(x) = \mathrm{Fl}^X_t \circ \mathrm{Fl}^X_s(x),$$

whenever both sides exist. Moreover, the mapping $\mathrm{Fl}^X_{t_0}(x)$ with a fixed parameter t is differentiable at all points x where it is defined.

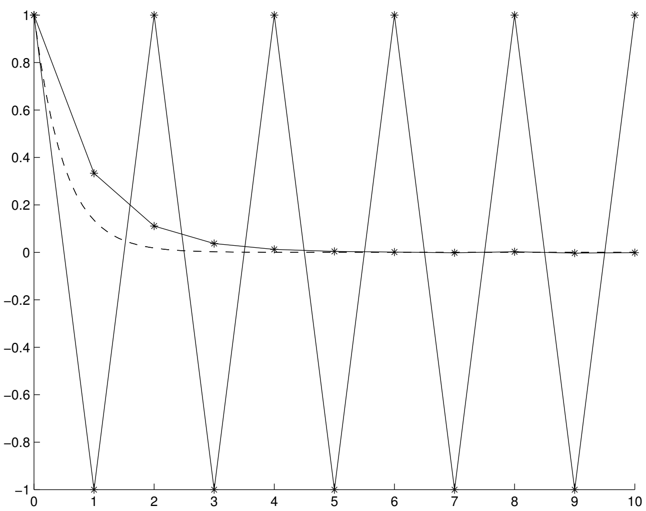
If a vector field X is defined on the whole $\mathbb{R}^n$ and has compact support, then its flow clearly exists at all points and for all values of t. Such vector fields are called *complete*. The flow of a complete vector field thus consists of diffeomorphisms $\mathrm{Fl}^X_t \, \mathbb{R}^n \to \mathbb{R}^n$ with inverse diffeomorphisms Fl^X_{-t}.

A simple example of a complete vector field is the field $X(x) = \frac{\partial}{\partial x_1}$. Its flow is given by

$$\mathrm{Fl}^X_t(x_1, \dots, x_n) = (x_1 + t, x_2, \dots, x_n).$$

On the other hand, the vector field $X(t) = t^2\frac{d}{dt}$ on the one-dimensional space $\mathbb{R}$ is not complete as its solutions are of the form

$$t \mapsto \frac{1}{C - t}$$



□

for initial conditions with $t_0 \neq 0$, so they "run away" towards infinite values in a finite time.

The description of a vector fields as assigning the tangent vector in the direction space to each point of the Euclidean space is independent of the coordinates. The following theorem thus gives us a geometric local qualitative description of all solutions of systems of ordinary differential equations in a neighborhood of each point x where the given vector field X is non-zero.

Theorem. *If X is a vector field defined on a neighborhood of a point $x_0 \in \mathbb{R}^n$ and we have $X(x_0) \neq 0$, 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. We will construct a diffeomorphism $F = (f_1, \dots, f_n)$ step by step. Geometrically, the essence of the proof can be summarized as follows: we select a hypersurface which is complementary to the directions $X(x)$, goes through the point x_0, then we fix the coordinates on it, and finally, we extend them to some neighborhood of the point x_0 using the flow of the field X.

First, we move the point x_0 to the origin and use a linear transformation on $\mathbb{R}^n$ in order to achieve $X(0) = \frac{\partial}{\partial x_1}(0)$. Now, let us write in these coordinates $(x_1, \dots, x_n)$ the flow of the field X going through the point $(x_1, \dots, x_n)$ at time $t = 0$ as $x_i(t) = \varphi_i(t, x_1, \dots, x_n)$. We define

$$f_i(x_1, \dots, x_n) = \varphi_i(x_1, 0, x_2, \dots, x_n).$$

Since the flow of the field X satisfies

$$\varphi_i(0, 0, x_2, \dots, x_n) = (0, x_2, \dots, x_n),$$

since it is the flow at time 0, we obtain

$$\frac{\partial F}{\partial x_i}(0) = (0, \dots, 1, \dots, 0), \quad i = 2, \dots, n,$$

and the same formula holds for $i = 1$, because we have $X = \frac{\partial}{\partial x_1}$. Therefore, the Jacobian matrix of the mapping F at the origin is the identity matrix E, so it is indeed a transformation of coordinates on some neighborhood (see the inverse mapping theorem in paragraph 8.17).

Now, directly from the definition of the mapping F in terms of the flow of the vector field X, the flow of the field will be expressed in the new coordinates $(y_1, \dots, y_n)$ as

$$\mathrm{Fl}_t^X(y_1, \dots, y_n) = (y_1 + t, y_2, \dots, y_n).$$

Verify this by yourselves in detail! □

8.56. Higher-order equations. An ordinary differential equation of order k (solved with respect to the highest derivative) is an equation

$$y^{(k)}(t) = f(t, y(t), y'(t), \dots, y^{(k-1)}(t)),$$

where f is a known function of $k+1$ variables, x is an independent variable, and $y(x)$ is an unknown function of one variable. We will show that this type of equation is always equivalent to a system of k first-order equations.

We introduce new unknown functions in a variable t as follows: $y_0(t) = y(t)$, $y_1(t) = y_0'(t)$, …, $y_{k-1}(t) = y_{k-2}'(t)$. Now, the function $y(t)$ is a solution of our original equation if and only if it is the first component of the system of equations

$$\begin{aligned} y_0'(t) &= y_1(t)\\ y_1'(t) &= y_2(t)\\ &\vdots\\ y_{n-2}'(t) &= y_{n-1}(t)\\ y_{n-1}'(t) &= f(t, y_0(t), y_1(t), \dots, y_{n-1}(t)). \end{aligned}$$

We thus get the following direct corollary of the theorems from 8.52–8.54:

Solutions of higher-order ODEs

Theorem. *Let a function $f(t, y_0, \dots, y_{k-1}) : U \subset \mathbb{R}^{k+1} \to \mathbb{R}$ have continuous partial derivatives on an open set U. Then, for every point $(t_0, z_0, \dots, z_{k-1}) \in U$, there exists a maximal interval $I_{max} = [x_0 - a, x_0 + b]$, with positive numbers $a, b \in \mathbb{R}$, and a unique function $y(t) : \mathbb{I}_{max} \to \mathbb{R}$ which is a solution of the k-th order equation*

$$y^{(k)}(t) = f(t, y(t), y'(t), \dots, y^{(k-1)}(t))$$

with initial condition

$$y(t_0) = z_0, y'(t_0) = z_1, \dots, y^{(k-1)}(t_0) = z_{k-1}.$$

Moreover, this solution depends differentiably on the initial condition and potential further parameters differentiably entering the function f.

We can thus see that for an unambiguous assignment of a solution of an ordinary k-th order differential equation, we have to determine at a point the value and the first $k - 1$ derivatives of the resulting function.

If we worked with a system of ℓ equations of order k, then 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 will hold again.

Of course, stronger properties pass on to all such systems in the cases when the right-hand side of the equation f is differentiable up to order k (inclusive) or analytic, including the parameters, and these properties pass on to the solutions as well.

8.57. Linear differential equations. We have already perceived the operation of differentiation as a linear mapping from (sufficiently) smooth functions to functions. If we multiply the derivatives $(\frac{d}{dx})^j$ of the particular orders j by fixed functions $a_j(t)$ and add up these expressions, we get the so-called *linear differential operator*:

$$y(t) \mapsto D(y)(t) = a_k(t)y^{(k)}(t) + \dots + a_1(t)y'(t) + a_0y(t).$$

To solve the corresponding *homogeneous linear differential equation* then means to find a function y satisfying $D(y) = 0$, i. e., the image is the identically zero function.

It is clear straight from the definition that the sum of two solutions will again be a solution, since for any functions y_1 and y_2, we have

$$D(y_1 + y_2)(t) = D(y_1)(t) + D(y_2)(t).$$

Analogously, a constant multiple of a solution is again a solution. The set of all solutions of a k-th order linear differential equation is thus a vector space. Applying the previous theorem about existence and uniqueness, we get the following:

The space of solutions of linear equations

Theorem. *The set of all solutions of a homogeneous linear differential equation of order k is always a vector space of dimension k. Therefore, we can always describe the solutions 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. If we choose k linearly independent initial conditions at a fixed point, then we get, for each of them, a unique solution of our solution. A linear combination of these initial condition then leads to the same linear combination of the corresponding solutions. We thus exhaust all of the possible initial conditions, so we get the entire space of solutions of our equation in this way. □

8.58. Linear differential equations with constant coefficients. The previous discussion surely reminded us the situation with homogeneous linear difference equations we dealt with in paragraph 3.9 of the third chapter. The analogy goes further even when all of the coefficients a_j of the differential operator D are constant. We have already seen such first-order equations (8.8) whose solution is an exponential with an appropriate constant at the argument. Just like in the case of difference equations, it suggests itself to try whether such a form of the solution $y(t) = \mathrm{e}^{\lambda t}$ with an unknown parameter λ can satisfy an equation of order k. Substitution yields

$$D(\mathrm{e}^{\lambda t}) = \big(a_k\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0(x)\big)\,\mathrm{e}^{\lambda t}\,.$$

The parameter λ thus leads to a solution of a linear differential equation with constant coefficients if and only if λ is a root of the so-called *characteristic polynomial* $a_k\lambda^k + \cdots + a_1\lambda + a_0$.

If this polynomial has k distinct roots, then we get 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 $y(t) = t\,\mathrm{e}^{\lambda t}$ is also a solution. Similarly, for higher multiplicities ℓ, we get ℓ distinct solutions $\mathrm{e}^{\lambda t}, t\,\mathrm{e}^{\lambda t}, \dots, t^{\ell}\,\mathrm{e}^{\lambda t}$.

In the case of a general linear differential equation, we assign a non-zero value of the differential operator D. Again, analogously to the reasonings about systems of linear equations or linear difference equations, we can see that the general solution of this type of (non-homogeneous) equation

$$D(y)(t) = b(t),$$

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 thus again a nice finite-dimensional affine space, hidden in a 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 upon looking for the solution in a similar form as the right-hand side is.

8.59. Matrix systems with constant coefficients. Now, let us take a look at a very special case of first-order systems, whose right-hand side is given by multiplication by an n^2-dimensional unknown vector function $Y(t)$:

$$Y'(t) = A \cdot Y(t) \tag{8.11}$$

with a constant matrix $A \in \mathrm{Mat}_n(\mathbb{R})$. Combining all our knowledge from linear algebra and univariate function analysis, we can guess the solution directly if we define the so-called *exponential of a matrix* by the formula

$$B(t) = \mathrm{e}^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k.$$

The right-hand expression can be formally viewed as a matrix whose entries b_{ij} are infinite series created from the mentioned products. If we bound all entries of A by the maximum of their absolute values $\|A\| = C$, then, for the k-th summand in $b_{ij}(t)$, we get the bound $\frac{t^k}{k!} n^k C^k$ 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}. Trying to differentiate the terms of our series one by one, we get a uniformly convergent series with limit $A\,\mathrm{e}^{tA}$. Therefore, by the general properties of uniformly convergent series, the derivative

$$\frac{d}{dt}\left(\mathrm{e}^{tA}\right) = A\,\mathrm{e}^{tA}$$

also equals this expression. We have thus obtained the general solution of our system (8.11) in the form

$$Y(t) = \mathrm{e}^{tA} \cdot Z,$$

where $Z \in \mathrm{Mat}_n(\mathbb{R})$ is an arbitrary constant matrix. Indeed, the exponential e^{tA} is an invertible matrix for all t, so we have obtained a vector space of the proper dimension, and hence all general solutions.

Noticeably, if we have only a vector equation with a constant matrix $A \in \mathrm{Mat}_n(\mathbb{R})$, $y'(t) = A \cdot y(t)$, for an unknown function $y : \mathbb{R} \to \mathbb{R}^n$, then the exponential e^{tA} gives n linearly independent solutions with its n columns. The general solution is then given by any linear combination of them.

Finally, let us recall that we met the first-order matrix system in paragraph 8.54 when we were thinking about the derivative of the solutions of vector equations with respect to the initial conditions. Now, consider a differentiable vector field $X(x)$ defined on a neighborhood of a point $x_0 \in \mathbb{R}^n$ such that $X(x_0) = 0$. Then, the point x_0 is a fixed point of its flow $\mathrm{Fl}_t^X(x)$.

The differential $\Phi(t) = D_x \mathrm{Fl}_t^X(x_0)$ satisfies (see (e8.42b) on page 531)

$$\Phi'(t) = D^1X(x_0) \cdot \Phi(t), \quad \Phi(0) = E.$$

We thus know explicitly the evolution of the differential of the vector field's flow at the singular point x_0, which is given by the exponential

$$\Phi(t) = \mathrm{e}^{tA}, \quad A = D^1X(x_0).$$

This is a useful step for reasoning about the qualitative behavior in a neighborhood of the stationary point x_0.

8.60. A note about Markov chains. In the third chapter, we dealt with iterative processes, where the so-called stochastic matrices and Markov processes determined by them played an interesting role. Let us recall that a matrix A is stochastic iff the sum of each of its columns is one. In other words,

$$(1\ \ldots 1) \cdot A = (1\ \ldots\ 1).$$

If we take the exponential e^{tA}, we obtain

$$(1\ \ldots 1) \cdot \mathrm{e}^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!}(1\ \ldots 1) \cdot A^k = \mathrm{e}^t (1\ \ldots 1).$$

Therefore, for every t, the invertible matrix $B(t) = \mathrm{e}^{-t}\,\mathrm{e}^{tA}$ is stochastic. We thus get a continuous version of the Markov process (infinitesimally) generated by the stochastic matrix A.

Indeed, differentiating with respect to t, we obtain

$$\frac{d}{dt}B(t) = -\,\mathrm{e}^{-t}\,\mathrm{e}^{tA} + \mathrm{e}^{-t}\,A\,\mathrm{e}^{tA} = (-E + A)B(t),$$

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 quite intuitively. If the matrix A is stochastic, then the instantaneous increase of the vector $y(t)$ in the vector system with the matrix A, $y'(t) = A \cdot y(t)$, is again a stochastic vector. However, we want the Markov process to keep the vector $y(t)$ stochastic for all t. Hence, the sum of increases of the particular components of the vector $y(t)$ must be zero, which is guaranteed by subtracting the identity matrix.

As we have seen above, the columns of the matrix solution $Y'(t)$ create a basis of all solutions $y'(t)$ of the vector system.

Let us further suppose that the matrix A is primitive, i. e., some of its powers has only positive entries, see **??** on page **??**. Then we know that its powers converge to a matrix A_∞, all of whose columns are eigenvectors corresponding to the eigenvalue 1. Hence, there must exist a universal constant bound for all powers $\|A^k - A_\infty\| \leq C$, and for every small positive ε, there is an $N \in \mathbb{N}$ such that for all $k \geq N$, we already have that $\|A^k - A_\infty\| \leq \varepsilon$. Now, we can bound the difference between the solution $Y'(t)$ for large t and the constant matrix A_∞:

$$\left\| \mathrm{e}^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k - \mathrm{e}^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A_\infty \right\|$$

$$\leq \mathrm{e}^{-t} \sum_{k<N} \frac{t^k}{k!} C \|A_\infty\| + \mathrm{e}^{-t}\,\varepsilon \|A_\infty\|.$$

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. Indeed, differentiation of the sum yields the same, only for N one less, and the derivative in the denominator is not changed, so the limit is zero. Therefore, for our chosen ε, we can find a T such that $f(t)$ would be less than ε for $t \geq T$. The whole expression has thus been bound (for $n \geq N$ and $t \geq T > 0$) by the number $\varepsilon(C + 1)\|A_\infty\|$.

We have thus proves a very interesting statement, which resembles the discrete version of Markov processes:

Continuous processes with a stochastic matrix

Theorem. *Every primitive stochastic matrix A gives 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*

$$Y(t) = \mathrm{e}^{-t}\, e^{tA}\,,$$

- *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 t,*
- *every stochastic solution converges for $t \to \infty$ to an eigenvector y_∞ of the matrix A corresponding to the eigenvalue 1 of the matrix A.*

4. Notes about numerical methods

Except for the simple equations, like the linear ones with constant coefficients, we seldom encounter analytically solvable equations in practice. Therefore, we usually need some techniques to approximate the solutions of the equations we are working with.

We have already thought of a similar idea anywhere we dealt with approximations (i. e., we would recommend to compare this to the earlier paragraphs about splines, Taylor polynomials, and Fourier series). With a bit of courage, we can consider difference and differential equations to be mutual approximations. In one direction, we replace differences with differentials (for example, in economical or population models), and vice versa.

We will stop for a while to look at replacing derivatives with differences. First, we introduce the usual notation for bounds on the errors.

Let us recall that having a function $f(x)$ in variable x, we say that it is, in a neighborhood of a limit point x_0 of its domain, of *order of magnitude* $O(\varphi(x))$ for a function $\varphi(x)$ iff there exists a neighborhood U of the point x_0 and a constant C such that

$$|f(x)| \le C \cdot |\varphi(x)|$$

for all $x \in U$. The limit point x_0 can also be one of the infinite values $\pm\infty$.

The most usual cases are $O(x^p)$ for a *polynomial order of magnitude*, at zero or infinity; $O(\ln x)$ for a *logarithmic order of magnitude* at infinity, and so on. Let us notice that the logarithmic order of magnitude is independent of the choice of the logarithm base.

A good example is the approximation of a function by its Taylor polynomial of order k at a point x_0. Taylor's theorem for univariate functions says that the error of this approximation is $O(h^{k+1})$, where h is the increase of the argument $x - x_0 = h$.

We also considered similar topics in the case of Fourier series.

8.61. Euler method. In the case of ordinary differential equations, the simplest scheme is approximation with the so-called Euler polygons. We will present it 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 upon time and their derivatives.

Let us consider an equation (of order one, for the sake of simplicity and without loss of generality)

$$y'(t) = f(t, y(t)).$$

We will denote the discrete increase of time by h, i. e., $t_n = t_0 + nh$, and $y_n = y(t_n)$. It follows from Taylor's theorem (with remainder of order two) and our equation that

$$y_{n+1} = y_n + y'(t_n)h + O(h^2) = y_n + f(t_n, y_n)h + O(h^2).$$

Therefore, if we make n such steps with increase h from t_0 to t_n, the expected bound for the total error following the local inaccuracies of our linear approximation at most $h\,O(h^2)$, i. e., the error's order of magnitude will be $O(h)$. In practice, rounding errors take effect as well.

Using the numerical procedure of Euler method, we consider the piecewise linear polygon defined above to be the solution.