Mathematics

Vladimíra Hájková, Michal Johanis,
Oldřich John, Ondřej F. K. Kalenda
and Miroslav Zelený

matfyzpress
PRAHA 2012

Contents

CHAPTER 1

Functions of several variables

In the previous chapter we studied functions of one real variable. However, it is usual that some quantity depends on more variables. That takes us to concept of function whose values depends of several real variables. In next sections we are going to deal at first with sets, which are domains of these functions, then we introduce basic notions of differential calculus of multivariate functions.

1.1. The set $\mathbb{R}^n$ as a metric and linear space

Let $n \in \mathbb{N}$. Remind that the set $\mathbb{R}^n$ consists of all n-tuples of real numbers, since it is a Cartesian product of n sets:

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{n\text{-times}}.$$

If $\vec{x} \in \mathbb{R}^n$, then we denote its i-th coordinate by x_i, and hence we can write $\vec{x} = [x_1, \ldots, x_n]$. There are some important elements in set $\mathbb{R}^n$. First of all it is **origin**, that is an element, whose all coordinates equals zero. We denote it by $\vec{o}$. For $i \in \{1, \ldots, n\}$ we define $\vec{e^i} \in \mathbb{R}^n$ as follows:

$$\vec{e^i} = [0, \ldots, 0, \underbrace{1}_{i\text{-th coordinate}}, 0, \ldots, 0].$$

These elements will be important to us further.

Elements of $\mathbb{R}^n$ can be added together and multiplicated by a real number: if $\vec{x} \in \mathbb{R}^n$, $\vec{x} = [x_1, \ldots, x_n]$, $\vec{y} \in \mathbb{R}^n$, $\vec{y} = [y_1, \ldots, y_n]$, $\lambda \in \mathbb{R}$, then we define

$$\vec{x} + \vec{y} = [x_1 + y_1, \ldots, x_n + y_n],$$
$$\lambda\vec{x} = [\lambda x_1, \ldots, \lambda x_n].$$

For operations addition and multiplication by a real number we introduced now, there are number of counting rules, derived from similar rules for real numbers (e.g. $\vec{x}+\vec{y} = \vec{y}+\vec{x}$, $\lambda(\vec{x}+\vec{y}) = \lambda\vec{x}+\lambda\vec{y}$). Think over that we can write every element $\vec{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$ in a form $\vec{x} = \sum_{i=1}^{n} x_i\vec{e^i}$.

The set $\mathbb{R}^n$ with operations addition and multiplication by a real number we will call the space $\mathbb{R}^n$ and the elements of $\mathbb{R}^n$ will be called the points of this space.

However, sometimes it is useful to look at a given element $\vec{x}$ from $\mathbb{R}^n$ as a *vector*, that means a directed line segment starting at the origin and ending in the point $\vec{x}$.

Now we introduce an important notion of distance.

Definition. The Euclidean metric (**distance**) on $\mathbb{R}^n$ is a function $\rho\colon \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty)$ defined by:

$$\rho(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

We call the number $\rho(\vec{x}, \vec{y})$ the **distance between points $\vec{x}$ and $\vec{y}$**.

Theorem 1 (properties of the Euclidean metric)**.** The Euclidean metric ρ has the following properties:

- $\forall \vec{x}, \vec{y} \in \mathbb{R}^n\colon \rho(\vec{x}, \vec{y}) = 0 \Leftrightarrow \vec{x} = \vec{y}$,
- $\forall \vec{x}, \vec{y} \in \mathbb{R}^n\colon \rho(\vec{x}, \vec{y}) = \rho(\vec{y}, \vec{x})$ (symmetry),
- $\forall \vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n\colon \rho(\vec{x}, \vec{z}) \leq \rho(\vec{x}, \vec{y}) + \rho(\vec{y}, \vec{z})$ (triangle inequality),
- $\forall \vec{x}, \vec{y} \in \mathbb{R}^n\ \forall \lambda \in \mathbb{R}\colon \rho(\lambda\vec{x}, \lambda\vec{y}) = |\lambda|\, \rho(\vec{x}, \vec{y})$ (homogeneity),
- $\forall \vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n\colon \rho(\vec{x} + \vec{z}, \vec{y} + \vec{z}) = \rho(\vec{x}, \vec{y})$ (translation invariance).

Proof. We prove only the triangle inequality. The other proofs are simple. Let $\vec{x} = [x_1, \ldots, x_n]$, $\vec{y} = [y_1, \ldots, y_n]$, $\vec{z} = [z_1, \ldots, z_n] \in \mathbb{R}^n$. We want to prove that:

$$\begin{aligned} \rho(\vec{x}, \vec{z}) = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2} \leq \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} + \sqrt{\sum_{i=1}^{n} (y_i - z_i)^2} = \\ = \rho(\vec{x}, \vec{y}) + \rho(\vec{y}, \vec{z}). \end{aligned} \tag{1}$$

We could write $a_i = x_i - y_i$, $b_i = y_i - z_i$ for $i = 1, \ldots, n$. Then (1) is equivalent to

$$\sqrt{\sum_{i=1}^{n} (a_i + b_i)^2} \leq \sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}. \tag{2}$$

Since all sums in (2) are non-negative, then (2) is equivalent to

$$\sum_{i=1}^{n} (a_i + b_i)^2 \leq \sum_{i=1}^{n} a_i^2 + 2\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2} + \sum_{i=1}^{n} b_i^2.$$

We alter this expression:

$$\sum_{i=1}^{n} a_i^2 + 2\sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} b_i^2 \leq \sum_{i=1}^{n} a_i^2 + 2\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2} + \sum_{i=1}^{n} b_i^2. \tag{3}$$

Firstly, inequality (3) corresponds to inequality (1). On the other hand, the inequality (3) follows from Cauchy's inequality (Example **??**). That completes the proof. ∎

The following notions, based on the distance definition, play the key role in the multivariate function theory.

Definition. Let $\vec{x} \in \mathbb{R}^n$, $r > 0$. We call $B(\vec{x}, r)$ defined by the expression

$$B(\vec{x}, r) = \{\vec{y} \in \mathbb{R}^n;\ \rho(\vec{x}, \vec{y}) < r\}$$

the **open ball** with centre $\vec{x}$ and radius r or the **neighbourhood** of point $\vec{x}$.

Definition. Let $\vec{x}^j \in \mathbb{R}^n$ for each $j \in \mathbb{N}$ and let $\vec{x} \in \mathbb{R}^n$. We say that a sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ **converges to $\vec{x}$** if $\lim\limits_{j\to\infty} \rho(\vec{x}, \vec{x}^j) = 0$. We call the element $\vec{x}$ **the limit of the sequence** $\{\vec{x}^j\}_{j=1}^{\infty}$. A sequence $\{\vec{y}^j\}_{j=1}^{\infty}$ of the elements of $\mathbb{R}^n$ is **convergent** if there exists $\vec{y} \in \mathbb{R}^n$ so that $\{\vec{y}^j\}_{j=1}^{\infty}$ converges to $\vec{y}$.

Remark. It follows directly from the definition that an element $\vec{x} \in \mathbb{R}^n$ is the limit of the sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ if and only if

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \exists j_0 \in \mathbb{N}\ \forall j \in \mathbb{N}, j \geq j_0 \colon \vec{x}^j \in B(\vec{x}, \varepsilon).$$

Compare with the definition of the limit of a sequence of real numbers on page **??**.

Theorem 2. Let $\vec{x}^j \in \mathbb{R}^n$ for each $j \in \mathbb{N}$ a $\vec{x} \in \mathbb{R}^n$. The sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ converges to $\vec{x}$ if and only if for each $i \in \{1, \ldots, n\}$ the number sequence $\{x_i^j\}_{j=1}^{\infty}$ converges to number x_i.

Proof. Suppose that sequence $\{\vec{x}^j\}$ converges to $\vec{x}$, according to the definition it means

$$\lim_{j\to\infty} \sqrt{\sum_{k=1}^{n} (x_k^j - x_k)^2} = 0. \tag{4}$$

Choose fixedly $i \in \{1, \ldots, n\}$. Then for arbitrary $j \in \mathbb{N}$ holds

$$\sqrt{\sum_{k=1}^{n} (x_k^j - x_k)^2} \geq \left|x_i^j - x_i\right| \geq 0. \tag{5}$$

From (4), (5) and the sandwich theorem (Theorem **??**) follows $\lim\limits_{j\to\infty} x_i^j = x_i$.

Now suppose that $\lim\limits_{j\to\infty} x_i^j = x_i$ for each $i \in \{1, \ldots, n\}$. Applying the theorem about limit arithmetic (Theorem **??**) and using the continuity of function $t \mapsto \sqrt{t}$ on interval $[0, +\infty)$ we deduce (4). ∎

Remark. Theorem 2 says that convergence in the space $\mathbb{R}^n$ is the same as convergence by coordinates. It also implies that a sequence $\{\vec{x}^j\}$ of elements of $\mathbb{R}^n$ has at most one limit. Therefore it is correct to denote the limit of a sequence $\{\vec{x}^j\}$ (if it exists) by the symbol $\lim\limits_{j\to\infty} \vec{x}^j$. We will sometimes write $\vec{x}^j \to \vec{x}$ instead of $\lim\limits_{j\to\infty} \vec{x}^j = \vec{x}$.

Definition.

(i) Let $M \subset \mathbb{R}^n$. We say that $\vec{x} \in \mathbb{R}^n$ is an **interior point of a set** M if there exists $r > 0$, such that $B(\vec{x}, r) \subset M$.

(ii) We call $M \subset \mathbb{R}^n$ **open in** $\mathbb{R}^n$, if all of its points are interior points.

(iii) A **interior of a set** M is the set of all interior points of M. We denote the interiorof the set M $\operatorname{Int} M$.

Example 3. Let $\vec{x} \in \mathbb{R}^n$ and $R > 0$. An open ball $B(\vec{x}, R)$ is an open set in $\mathbb{R}^n$.

Proof. We have to prove that each point of the set $B(\vec{x}, R)$ is its interior point. Let $\vec{y} \in B(\vec{x}, R)$. We want to find $r > 0$, such that $B(\vec{y}, r) \subset B(\vec{x}, R)$. Choose $r = R - \rho(\vec{x}, \vec{y})$. The number r is positive, because $\rho(\vec{x}, \vec{y}) < R$. Provided that $\vec{z} \in B(\vec{y}, r)$, we could use the triangle inequality

$$\rho(\vec{x}, \vec{z}) \le \rho(\vec{x}, \vec{y}) + \rho(\vec{y}, \vec{z}) < \rho(\vec{x}, \vec{y}) + r = R,$$

and then $\vec{z} \in B(\vec{x}, R)$. We have proved that $B(\vec{y}, r) \subset B(\vec{x}, R)$, and then the point $\vec{y}$ is an interior point of the set $B(\vec{x}, R)$. Draw a figure for $n = 2$. ∎

Now we provide some basic properties of open sets.

Theorem 4 (properties of open sets)**.**

(i) The empty set and the whole space $\mathbb{R}^n$ are open in $\mathbb{R}^n$.

(ii) Let A is an non-empty set of indexes. Let the sets $G_\alpha \subset \mathbb{R}^n$, $\alpha \in A$, are open in $\mathbb{R}^n$. Then $\bigcup_{\alpha\in A} G_\alpha$ is open set in $\mathbb{R}^n$.

(iii) Let $m \in \mathbb{N}$. Let sets G_i, $i = 1, \dots, m$, are open in $\mathbb{R}^n$. Then $\bigcap_{i=1}^m G_i$ is an open set in $\mathbb{R}^n$.[1]

Proof. (i) This proposition is obvious.

(ii) If $\vec{x} \in \bigcup_{\alpha\in A} G_\alpha$, then $\alpha_0 \in A$ can be found such that $\vec{x} \in G_{\alpha_0}$. Due to openness of the set G_{α_0} there exists $r > 0$ satisfying $B(\vec{x}, r) \subset G_{\alpha_0}$, and thus $B(\vec{x}, r) \subset \bigcup_{\alpha\in A} G_\alpha$. It means that $\vec{x}$ is an interior point of the set $\bigcup_{\alpha\in A} G_\alpha$ and the proposition is thus proved.

(iii) If $\vec{x} \in \bigcap_{i=1}^m G_i$, then for each $i \in \{1, \dots, m\}$ there exists $r_i > 0$ such as $B(\vec{x}, r_i) \subset G_i$, because G_i is open. Put $r = \min\{r_1, \dots, r_m\}$. Then we get $r > 0$ and $B(\vec{x}, r) \subset \bigcap_{i=1}^m G_i$. ∎

[1] The symbol $\bigcap_{i=1}^m$ means the same as $\bigcap_{i\in\{1,\dots,m\}}$.

Definition.

(i) Let $M \subset \mathbb{R}^n$ and $\vec{x} \in \mathbb{R}^n$. We say that $\vec{x}$ is a **a limit point of the set** M, provided that for each $r > 0$ it is true that

$$B(\vec{x}, r) \cap M \neq \emptyset \quad \& \quad B(\vec{x}, r) \cap (\mathbb{R}^n \setminus M) \neq \emptyset.$$

(ii) The set of all limit points of M is called a **boundary** of M. We denote it by $H(M)$.

(iii) A **closure of the set** M is the set $M \cup H(M)$. We denote the closure of the set M by $\overline{M}$.

(iv) We say that the set M is **closed in** $\mathbb{R}^n$ if it contains all of its limit points (i.e. $H(M) \subset M$ or $\overline{M} = M$).

Theorem 5 (Characterization of closed sets)**.** Let $M \subset \mathbb{R}^n$. Then the following conditions are equivalent.

(i) The set M is closed in $\mathbb{R}^n$.

(ii) The set $\mathbb{R}^n \setminus M$ is open in $\mathbb{R}^n$.

(iii) If $\vec{x} \in \mathbb{R}^n$ is the limit of a convergent sequence $\{\vec{x}^j\}$ of points of the set M, then $\vec{x} \in M$.

Proof. The structure of the proof will be as follows: First, we shall prove that the condition (i) imply (ii), than we shall prove the implication (ii) $\Rightarrow$ (iii) and in the third step we shall prove from (iii) the condition (i). This would complete the proof of the theorem is what had to be proved, because from proved implications follow the rest.

(i) $\Rightarrow$ (ii) Let $\vec{x} \in \mathbb{R}^n \setminus M$. From the assumption that the set M is closed, we get $x \notin H(M)$. It implies that there exists $r > 0$ such that $B(\vec{x}, r) \cap M = \emptyset$ or $B(\vec{x}, r) \cap (\mathbb{R}^n \setminus M) = \emptyset$. The second eventuality could not happen in our case, because $\vec{x} \in \mathbb{R}^n \setminus M$. Hence $B(\vec{x}, r) \cap M = \emptyset$, in other words $B(\vec{x}, r) \subset \mathbb{R}^n \setminus M$, and thus $\vec{x}$ is an interior point of $\mathbb{R}^n \setminus M$.

(ii) $\Rightarrow$ (iii) Consider the sequence $\{\vec{x}^j\}$ of points of the set M which converges to the element $\vec{x} \in \mathbb{R}^n$. For each $r > 0$ there exists $j \in \mathbb{N}$ such that $\vec{x}^j \in B(\vec{x}, r)$. It means that $\vec{x}$ is not an inner point of $\mathbb{R}^n \setminus M$. The set $\mathbb{R}^n \setminus M$ is open and thus contains only its inner points, so $\vec{x} \notin \mathbb{R}^n \setminus M$, i.e. $\vec{x} \in M$.

(iii) $\Rightarrow$ (i) We assume that (iii) holds and we want to deduce $H(M) \subset M$. Let $\vec{x} \in H(M)$. Then for each $j \in \mathbb{N}$ we obtain $B(\vec{x}, 1/j) \cap M \neq \emptyset$. Thus there exists $\vec{x}^j \in B(\vec{x}, 1/j) \cap M$ for each $j \in \mathbb{N}$. Then $\lim \vec{x}^j = \vec{x}$, because $0 \leq \rho(\vec{x}, \vec{x}^j) \leq 1/j$, $j \in \mathbb{N}$. According to (iii) we obtain $\vec{x} \in M$ and this is what had to be proved. ∎

Theorem 6 (properties of closed sets)**.**

(i) The empty set and the whole space $\mathbb{R}^n$ are closed in $\mathbb{R}^n$.

(ii) Let A be an non-empty set of indexes. Let the sets $F_\alpha \subset \mathbb{R}^n$, $\alpha \in A$, be closed in $\mathbb{R}^n$. Then $\bigcap_{\alpha\in A} F_\alpha$ is a closed set in $\mathbb{R}^n$.
(iii) Let $m \in \mathbb{N}$. Let sets F_i, $i = 1, \ldots, m$, be closed in $\mathbb{R}^n$. Then $\bigcup_{i=1}^m F_i$ is a closed set in $\mathbb{R}^n$.[2]

Proof. All propositions could be easily proved from Theorem 4, the equivalence of (i) and (ii) in Theorem 5 and De Morgan's laws (Theorem **??**). ∎

In the following theorem we will deduce important properties of closure and interior of a set.

Theorem 7. Let $M \subset \mathbb{R}^n$. Then:
(i) The set $\overline{M}$ is closed in $\mathbb{R}^n$.
(ii) The set $\operatorname{Int} M$ is open in $\mathbb{R}^n$.
(iii) The set M is open in $\mathbb{R}^n$ if and only if $M = \operatorname{Int} M$.

Proof. (i) Suppose that $\vec{x} \in \mathbb{R}^n \setminus \overline{M}$. Then $\vec{x} \notin H(M)$ and thus there exists $\delta \in \mathbb{R}$, $\delta > 0$, such that $B(\vec{x}, \delta) \cap M = \emptyset$ or $B(\vec{x}, \delta) \cap (\mathbb{R}^n \setminus M) = \emptyset$. Since $\vec{x} \notin M$, the second eventuality could not happen. For each point $\vec{y} \in B(\vec{x}, \delta)$ there exists $\eta \in \mathbb{R}$, $\eta > 0$ such that $B(\vec{y}, \eta) \subset B(\vec{x}, \delta)$ and thus $B(\vec{y}, \eta) \cap M = \emptyset$. It implies $\vec{y} \notin H(M)$. Hence $B(\vec{x}, \delta) \cap H(M) = \emptyset$. This yields $B(\vec{x}, \delta) \cap \overline{M} = \emptyset$ and thus $\mathbb{R}^n \setminus \overline{M}$ is open in $\mathbb{R}^n$. According to the theorem 5 is then $\overline{M}$ closed in $\mathbb{R}^n$.

(ii) Suppose that $\vec{x} \in \operatorname{Int} M$. Thus, there exists $\delta \in \mathbb{R}$, $\delta > 0$ such that $B(\vec{x}, \delta) \subset M$. Then for arbitrary $\vec{y} \in B(\vec{x}, \delta)$ there exists $\eta \in \mathbb{R}$, $\eta > 0$ such that $B(\vec{y}, \eta) \subset B(\vec{x}, \delta) \subset M$. The point $\vec{y}$ is then an interior point of M. From that we obtain $B(\vec{x}, \delta) \subset \operatorname{Int} M$, what we wanted.

(iii) If M is open, then each of its points is an interior point of M and thus $M = \operatorname{Int} M$. The implicational converse follows directly from the definition of an open set. ∎

Remark. Remark that $\operatorname{Int} M$ is the biggest open set contained in M in the following sense: If G is a set open in $\mathbb{R}^n$ satisfying $G \subset M$, then $G \subset \operatorname{Int} M$. Similarly $\overline{M}$ is the smallest closed set containing M.

Definition. We say that a **set** $M \subset \mathbb{R}^n$ is **bounded** if there exists $r > 0$ satisfying $M \subset B(\vec{o}, r)$. A **sequence** of the points of $\mathbb{R}^n$ is **bounded** provided that the set of its terms is bounded.

Theorem 8. A set $M \subset \mathbb{R}^n$ is bounded if and only if the set $\overline{M}$ is bounded.

Proof. Suppose that the set M is bounded. Thus, there exists $r > 0$ such that $M \subset B(\vec{o}, r)$. If $\vec{x} \in \overline{M}$, then there exists $\vec{y} \in M$ such that $\rho(\vec{x}, \vec{y}) < 1$. If $\vec{x} \in M$, we could choose $\vec{y} = \vec{x}$, on the other hand if $\vec{x} \notin M$, then $\vec{x} \in H(M)$ and $\vec{y}$ could

[2] The symbol $\bigcup_{i=1}^m$ means the same as $\bigcup_{i\in\{1,\ldots,m\}}$.

be an arbitrary point of the set $B(\vec{x}, 1) \cap M$, which has to be non-empty. From the triagle inequality we get

$$\rho(\vec{o}, \vec{x}) \leq \rho(\vec{o}, \vec{y}) + \rho(\vec{y}, \vec{x}) \leq \rho(\vec{o}, \vec{y}) + 1 < r + 1.$$

Thus $\overline{M} \subset B(\vec{o}, r + 1)$ holds and then $\overline{M}$ is bounded.

Thus also M is bounded, because $M \subset \overline{M}$. ■

Example 9. Let $M = \{[x, y] \in \mathbb{R}^2;\ x > 0, y \geq 0\}$. Decide if the given set is open or closed, determine its boundary, closure and interior.

Solution. It can be easily proved that $B([x, y], \min\{x, y\}) \subset M$. Hence, if $[x, y] \in (0, \infty) \times (0, \infty)$, it is an interior point of the set. Neighbourhoods $B([x, y], |x|/2)$, where $x < 0$, and $B([x, y], |y|/2)$, where $y < 0$, are contained in the complement of the set M. A neighbourhood in a form $B([0, y], r)$, where $r > 0$, $y \geq 0$, always intersect both the set M and its complement, because for example $[r/2, y + r/2] \in M \cap B([0, y], r)$ and $[-r/2, y] \in (\mathbb{R}^2 \setminus M) \cap B([0, y], r)$. Similarly we could show, that a neighbourhood of the form $B([x, 0], r)$, where $r > 0$, $x \geq 0$, always intersects both M and its complement.

Hence, we get

$$\begin{aligned} \operatorname{Int} M &= \{[x, y] \in \mathbb{R}^2;\ x > 0, y > 0\}, \\ H(M) &= \{[0, y] \in \mathbb{R}^2;\ y \geq 0\} \cup \{[x, 0] \in \mathbb{R}^2;\ x \geq 0\} \text{ a} \\ \overline{M} &= \{[x, y] \in \mathbb{R}^2;\ x \geq 0, y \geq 0\}. \end{aligned}$$

It can be seen that $M \neq \operatorname{Int} M$ a $M \neq \overline{M}$. That is, the set M is neither open nor closed. ♣

Example 10. For each $k \in \mathbb{N}$ let a set be defined by

$$M_k = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 < (1 + 1/k)^2\}.$$

Determine a set $\bigcap_{k=1}^{\infty} M_k$.[3] Decide if the sets M_k, $\bigcap_{k=1}^{\infty} M_k$ are open or closed.

Solution. Each of the sets M_k is an open disc with radius $1 + 1/k$ and for each $k \in \mathbb{N}$ $M_{k+1} \subset M_k$ holds. It seems that $\bigcap_{k=1}^{\infty} M_k = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \leq 1\}$. Now we try to prove this conjecture.

Let us set $M = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \leq 1\}$. First, prove inclusion $M \subset \bigcap_{k=1}^{\infty} M_k$. If $[x, y] \in M$, then $x^2 + y^2 \leq 1 < 1 + 1/k$ for each $k \in \mathbb{N}$ and thus $[x, y] \in \bigcap_{k=1}^{\infty} M_k$.

Next, $\bigcap_{k=1}^{\infty} M_k \subset M$ holds. Since if $[x, y]$ is contained in each set M_k, is $x^2 + y^2 < 1 + 1/k$ for each $k \in \mathbb{N}$ and it follows that $x^2 + y^2 \leq \inf\{1 + 1/k;\ k \in \mathbb{N}\} = 1$ and $[x, y] \in M$.

[3] The symbol $\bigcap_{k=1}^{\infty}$ means the same as $\bigcap_{k \in \mathbb{N}}$.

The set M_k, $k \in \mathbb{N}$, is an open ball and hence it is open; M_k is not closed, because e.g. the point $[1 + 1/k, 0] \in \overline{M}_k \setminus M_k$. It is not difficult to think over that $H(M) = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 = 1\}$ and the set M is thus closed. However, it is not open, because e.g. the point $[0, 1] \in M \setminus \operatorname{Int} M$.

The given example shows that the intersect of infinitely many open sets does not have to be an open set. Compare with the theorem 4. ♣

1.2. Continuous functions of several variables

In this section we will show how the notion of continuity could be defined for **functions of several variables**, i.e. for functions of the type $f\colon M \to \mathbb{R}$, $M \subset \mathbb{R}^n$. We also introduce some basic properties of this notion.

Definition. Let $M \subset \mathbb{R}^n$, $\vec{x} \in M$ and f be a function of n variables. We say that f is **continuous at a point $\vec{x}$ in M**, provided that

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \exists \delta \in \mathbb{R}, \delta > 0\ \forall \vec{y} \in B(\vec{x}, \delta) \cap M\colon f(\vec{y}) \in B(f(\vec{x}), \varepsilon).$$

To say that f is **continuous at** $\vec{x}$, means that f is continuous at $\vec{x}$ in some neighbourhood of the point $\vec{x}$, i.e.

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \exists \delta \in \mathbb{R}, \delta > 0\ \forall \vec{y} \in B(\vec{x}, \delta)\colon f(\vec{y}) \in B(f(\vec{x}), \varepsilon).$$

We state two theorems which shows behaviour of the defined notion while using arithmetic operations and function composition. They can be proved by a modification of the proofs of the theorem about function limit arithmetic (Theorem **??**) and the theorem about a limit of a composite function (Theorem **??**) respectively. We omit theese proofs.

Theorem 11. Let $M \subset \mathbb{R}^n$, $\vec{x} \in M$, $f\colon M \to \mathbb{R}$, $g\colon M \to \mathbb{R}$ and $c \in \mathbb{R}$. If f and g are continuous at a point $\vec{x}$ in M, then so are functions cf, $f + g$ and fg. If above that $g(\vec{x}) \neq 0$ holds, then also function f/g is continuous at a point $\vec{x}$ in M.

Theorem 12. Let $r, s \in \mathbb{N}$ and let $M \subset \mathbb{R}^s$, $L \subset \mathbb{R}^r$ and $\tilde{\vec{x}} \in M$. Let $\varphi_1, \ldots, \varphi_r$ are functions defined on M, continuous at the point $\tilde{\vec{x}}$ in M and $[\varphi_1(\vec{x}), \ldots, \varphi_r(\vec{x})] \in L$ for each $\vec{x} \in M$. Let $f\colon L \to \mathbb{R}$ is continuous at the point $[\varphi_1(\tilde{\vec{x}}), \ldots, \varphi_r(\tilde{\vec{x}})]$ in L. Then the composite function $F\colon M \to \mathbb{R}$ defined by

$$F(\vec{x}) = f\big(\varphi_1(\vec{x}), \varphi_2(\vec{x}), \ldots, \varphi_r(\vec{x})\big), \quad \vec{x} \in M,$$

is continuous at $\tilde{\vec{x}}$ in M.

The connection between continuity of a function and convergence of a sequence is stated in already mentioned Heine theorem.

Theorem 13 (Heine theorem)**.** Let $M \subset \mathbb{R}^n$, $\vec{x} \in M$ and $f\colon M \to \mathbb{R}$. Then the following conditions are equivalent:

(i) f is continuous at $\vec{x}$ in M,

(ii) $\lim\limits_{j\to\infty} f(\vec{x}^j) = f(\vec{x})$ for every sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ provided $\vec{x}^j \in M$ for $j \in \mathbb{N}$ and $\lim\limits_{j\to\infty} \vec{x}^j = \vec{x}$.

Proof. (i) $\Rightarrow$ (ii) Choose an arbitrary sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ of points of the set M, which converges to $\vec{x}$. Choose $\varepsilon > 0$. From the continuity of f at the point $\vec{x}$ in M follows the existence of $\delta > 0$ such that $f(\vec{y}) \in B(f(\vec{x}), \varepsilon)$ holds for each $\vec{y} \in B(\vec{x}, \delta) \cap M$. For that $\delta > 0$ we can find $j_0 \in \mathbb{N}$ satisfying $\rho(\vec{x}^j, \vec{x}) < \delta$ for $j \geq j_0$. If thus $j \geq j_0$, then $\vec{x}^j \in B(\vec{x}, \delta) \cap M$ and thus $f(\vec{x}^j) \in B(f(\vec{x}), \varepsilon)$.

(ii) $\Rightarrow$ (i) Prove non (i) $\Rightarrow$ non (ii). Suppose (i) does not hold. That is

$$\exists \varepsilon \in \mathbb{R}, \varepsilon > 0\ \forall \delta \in \mathbb{R}, \delta > 0\ \exists \vec{y} \in B(\vec{x}, \delta) \cap M\colon f(\vec{y}) \notin B(f(\vec{x}), \varepsilon).$$

For each $j \in \mathbb{N}$ we can find a point $\vec{y}^j \in B(\vec{x}, 1/j) \cap M$ satisfying $f(\vec{y}^j) \notin B(f(\vec{x}), \varepsilon)$. We have $\vec{y}^j \to \vec{x}$, but the sequence $\{f(\vec{y}^j)\}_{j=1}^{\infty}$ does not converge to $f(\vec{x})$, and thus (ii) does not hold. ∎

Definition. Let $M \subset \mathbb{R}^n$ and $f\colon M \to \mathbb{R}$. We say that f is **continuous on a set** M if and only if it is continuous at each point $\vec{x} \in M$ in M.

Remark. The definition is consistent with the previously defined notion continuity on an interval.

Example 14. Let $i, n \in \mathbb{N}$, $i \leq n$. A function $\pi^i\colon \mathbb{R}^n \to \mathbb{R}$ defined by $\pi^i(x_1, \ldots, x_n) = x_i$ is continuous on $\mathbb{R}^n$. We call these functions **coordinate projections**.

Proof. Prove that π^i is continuous at an arbitrary point $\tilde{\vec{x}} = [\tilde{x}_1, \ldots, \tilde{x}_n] \in \mathbb{R}^n$. Let $\varepsilon > 0$ be given, then set $\delta = \varepsilon$. Then for each $\vec{x} \in B(\tilde{\vec{x}}, \delta)$ it follows that

$$\left|\pi^i(\vec{x}) - \pi^i(\tilde{\vec{x}})\right| = |x_i - \tilde{x}_i| \leq \rho(\vec{x}, \tilde{\vec{x}}) < \delta = \varepsilon.$$

∎

Example 15. The function $\vec{x} \mapsto \rho(\vec{x}, \vec{o})$ is continuous on $\mathbb{R}^n$.

Proof. The following inequalities hold for each $\vec{u}, \vec{v} \in \mathbb{R}^n$:

$$\rho(\vec{u}, \vec{o}) \leq \rho(\vec{u}, \vec{v}) + \rho(\vec{v}, \vec{o}) \quad \text{a} \quad \rho(\vec{v}, \vec{o}) \leq \rho(\vec{v}, \vec{u}) + \rho(\vec{u}, \vec{o}),$$

from that it easily follows

$$|\rho(\vec{u}, \vec{o}) - \rho(\vec{v}, \vec{o})| \leq \rho(\vec{u}, \vec{v}).$$

Choose now a point $\vec{x} \in \mathbb{R}^n$ fixedly and let $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$ be given. Then for each $\vec{y} \in B(\vec{x}, \varepsilon)$ from the previous inequality it follows that

$$|\rho(\vec{y}, \vec{o}) - \rho(\vec{x}, \vec{o})| \leq \rho(\vec{y}, \vec{x}) < \varepsilon,$$

and the continuity of the given function at the point $\vec{x}$ is proved. ∎

Theorem 16. Let f be a continuous function on $\mathbb{R}^n$ and $c \in \mathbb{R}$. Then:

(i) The set $\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) < c\}$ is an open set in $\mathbb{R}^n$.
(ii) The set $\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) > c\}$ is an open set in $\mathbb{R}^n$.
(iii) The set $\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) \leq c\}$ is a closed set $\mathbb{R}^n$.
(iv) The set $\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) \geq c\}$ is a closed set $\mathbb{R}^n$.
(v) The set $\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) = c\}$ is a closed set $\mathbb{R}^n$.

Proof. (i) Let us set $M = \{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) < c\}$. If $\tilde{\vec{x}} \in M$, then we want to prove that we can find a neighbourhood $B(\tilde{\vec{x}}, r)$ which is a subset of M. Since $f(\tilde{\vec{x}}) < c$, we can set $\varepsilon = c - f(\tilde{\vec{x}})$ and notice that $\varepsilon > 0$. From the continuity of the function f at the point $\tilde{\vec{x}}$ it follows that we can find $r > 0$ such that

$$\forall \vec{x} \in B(\tilde{\vec{x}}, r)\colon\ f(\tilde{\vec{x}}) - \varepsilon < f(\vec{x}) < f(\tilde{\vec{x}}) + \varepsilon. \tag{6}$$

From the definition of the number ε and from (6) follows

$$\forall \vec{x} \in B(\tilde{\vec{x}}, r)\colon\ f(\vec{x}) < c,$$

and thus $B(\tilde{\vec{x}}, r) \subset M$.

Statement (ii) can be proved similarly or we can use the function $-f$. Statement (iii) follows from (ii) and the theorem 5, statement (iv) follows from (i) and Theorem 5. It remains to prove the statement (v). It follows that

$$\{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) = c\} = \{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) \leq c\} \cap \{\vec{x} \in \mathbb{R}^n;\ f(\vec{x}) \geq c\}.$$

This relation, (iii), (iv) and the theorem 6 imply (v). ∎

Definition. We call a set $M \subset \mathbb{R}^n$ **compact** if and only if every sequence of elements of the set M has a convergent subsequence whose limit is in M.[4]

The following theorem gives an important characterization of the compact subsets of $\mathbb{R}^n$, which we will use in finding extremes of a multivariate function.

Theorem 17 (characterization of compact sets in $\mathbb{R}^n$)**.** A set $M \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.

In the proof we will use the following lemma.

Lemma 18. Let $\{\vec{x}^j\}_{j=1}^{\infty}$ be a bounded sequence in $\mathbb{R}^n$. Then it has a convergent subsequence.

Proof. We prove the result by applying mathematical induction on n. For $n = 1$ it is the Bolzano-Weierstraß theorem (Theorem **??**).

Suppose that the statement holds for every bounded sequence in $\mathbb{R}^n$. Let $\{\vec{x}^j\}_{j=1}^{\infty}$ be a bounded sequence in $\mathbb{R}^{n+1}$, that means we can find $R > 0$ such that $\vec{x}^j \in$

[4] We will define subsequence of a sequence of elements of $\mathbb{R}^n$ in a similar way to defining subsequence of the real numbers.

$B(0,R)$ for each $j \in \mathbb{N}$. Denote $\vec{y}^j = [x_1^j, \ldots, x_n^j] \in \mathbb{R}^n$, $j \in \mathbb{N}$. Then $\rho(0,\vec{y}^j) \leq \rho(0,\vec{x}^j) < R$ holds for each $j \in \mathbb{N}$, and the sequence $\{\vec{y}^j\}_{j=1}^{\infty}$ is thus bounded. According to the induction assumption, the sequence $\{\vec{y}^j\}_{j=1}^{\infty}$ has a convergent subsequence $\{\vec{y}^{j_k}\}_{k=1}^{\infty}$. Next $\left|x_{n+1}^{j_k}\right| \leq \rho(0,\vec{x}^{j_k}) < R$ holds for each $k \in \mathbb{N}$, and the sequence of real numbers $\{x_{n+1}^{j_k}\}_{k=1}^{\infty}$ is thus bounded. According to the Bolzano-Weierstraß theorem, it has a convergent sebsequence $\{x_{n+1}^{j_{k_i}}\}_{i=1}^{\infty}$.

From the Theorem 2 follows that the sequence of real numbers $\{y_l^{j_k}\}_{k=1}^{\infty}$ is convergent for each $l \in \{1,\ldots,n\}$. From the theorem about a limit of a subsequence (Theorem **??**) follows that the subsequence $\{y_l^{j_{k_i}}\}_{i=1}^{\infty}$ is convergent. $\vec{x}^{j_{k_i}} = \left[y_1^{j_{k_i}}, \ldots, y_n^{j_{k_i}}, x_{n+1}^{j_{k_i}}\right]$ holds, and therefore according to Theorem 2 the sequence $\{\vec{x}^{j_{k_i}}\}_{i=1}^{\infty}$ is convergent. ∎

Proof of the Theorem 17. $\Rightarrow$ Let M be compact and not bounded. Then for each $j \in \mathbb{N}$ there exists $\vec{x}^j \in M \setminus B(\vec{o},j)$. But the sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ has a subsequence $\{\vec{x}^{j_k}\}_{k=1}^{\infty}$ which converges to a limit $\vec{y} \in M$. Then

$$j_k \leq \rho(\vec{x}^{j_k},\vec{o}) \leq \rho(\vec{x}^{j_k},\vec{y}) + \rho(\vec{y},\vec{o}).$$

We get $\lim\limits_{k\to\infty} j_k = +\infty$ and concurrently $\lim\limits_{k\to\infty} \big(\rho(\vec{x}^{j_k},\vec{y}) + \rho(\vec{y},\vec{o})\big) = \rho(\vec{y},\vec{o}) \in \mathbb{R}$. With the Theorem **??** we get a contradiction.

Show the closeness of the set M. Suppose that $\{\vec{x}^j\}_{j=1}^{\infty}$ is convergent sequence of the elements of the set M. We denote the limit of this sequence by $\vec{x}$. The set M is compact, and thus $\{\vec{x}^j\}_{j=1}^{\infty}$ has a subsequence, which converges to a limit in M. However, this limit must equal to $\vec{x}$ (Theorem **??** and Theorem 2), thus $\vec{x} \in M$. Closeness of the set M now follows from the Theorem 5.

$\Leftarrow$ Let $M \subset \mathbb{R}^n$ be bounded and closed set. Take an arbitrary sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ of the elements of the set M. This sequence is bounded, according Lemma 18 it has a subsequence $\{\vec{x}^{j_k}\}_{k=1}^{\infty}$, which converges to any $\vec{x} \in \mathbb{R}^n$. From the closeness of the set M follows according to the Theorem 5, that $\vec{x} \in M$. This is what had to be proved. ∎

Remark. From the previous theorem follows that:

- closed intervals in $\mathbb{R}$ are compact,
- finite unions of the closed intervals in $\mathbb{R}$ are compact,
- interval $(0,1)$ is not compact in $\mathbb{R}$.

Definition. Let $M \subset \mathbb{R}^n$, $\vec{x} \in M$ and f be a function defined at least on M (i.e. $M \subset D_f$).

- We say that f has a **maximum** (**minimum**, respectively) at a point $\vec{x}$ **on** $\boldsymbol{M}$ provided that

$$\forall \vec{y} \in M\colon f(\vec{y}) \leq f(\vec{x}) \qquad (\forall \vec{y} \in M\colon f(\vec{y}) \geq f(\vec{x}), \text{ respectively}).$$

We call the point $\vec{x}$ **maximum point** (**minimum point**, respectively) of the function f on M.

• We say that f has a **local maximum** (**local minimum**, respectively) at a point $\vec{x}$ **on** $\boldsymbol{M}$, if there exists $\delta > 0$ such that

$$\forall \vec{y} \in B(\vec{x}, \delta) \cap M \colon f(\vec{y}) \leq f(\vec{x}) \quad (\forall \vec{y} \in B(\vec{x}, \delta) \cap M \colon f(\vec{y}) \geq f(\vec{x}) \text{ respectively}).$$

We call the point $\vec{x}$ a **local maximum point** (**local minimum point**, respectively) of the function f on a set M.

• We say that f has a **strict local maximum** (**strict local minimum**, respectively) at a point $\vec{x}$ **on** $\boldsymbol{M}$, if there exists $\delta > 0$ such that

$$\forall \vec{y} \in (B(\vec{x}, \delta) \setminus \{\vec{x}\}) \cap M \colon f(\vec{y}) < f(\vec{x})$$
$$(\text{resp. } \forall \vec{y} \in (B(\vec{x}, \delta) \setminus \{\vec{x}\}) \cap M \colon f(\vec{y}) > f(\vec{x})).$$

We call the point $\vec{x}$ a **strict local maximum point** (**strict local minimum point**, respectively) of the function f on the set M.

• We denote the biggest (smallest, respectively) value of the function f on M (provided that this value exists) by the symbol $\max_M f$ ($\min_M f$ respectively).

Remark. Speaking about a local extrem of the multivariate function (without mentioning the set) means a local extrem on some neighbourhood.

Theorem 19. (about having extremes) Let $M \subset \mathbb{R}^n$ be a non-empty compact set and $f \colon M \to \mathbb{R}$ be continuous on M. Then f has both maximum and minimum on M.

Proof. Denote $G = \sup f(M)$. According to Lemma **??** there exists a sequence $\{y_j\}$ of elements of the set $f(M)$ such that $\lim y_j = G$. For each $j \in \mathbb{N}$ we can find $\vec{x}^j \in M$ satisfying $f(\vec{x}^j) = y_j$. The set M is compact, thus the sequence $\{\vec{x}^j\}_{j=1}^{\infty}$ has a subsequence $\{\vec{x}^{j_k}\}_{k=1}^{\infty}$, which cinverges to a limit $\vec{x}^* \in M$. The function f is continuous at the point $\vec{x}^*$ in M and thus according to Heine theorem (Theorem 13) $\lim\limits_{k\to\infty} f(\vec{x}^{j_k}) = f(\vec{x}^*)$ holds. On the other hand we have $\lim\limits_{k\to\infty} f(\vec{x}^{j_k}) = G$. Then $f(\vec{x}^*) = G$. This had proved that f has a maximum on M.

We could prove the existence of minimum similarly, or we can use a function $-f$, as in the proof of the Theorem **??**. ∎

From the previous theorem immediately follows the next corollary.

Corollary 20. Let $M \subset \mathbb{R}^n$ be a compact set and $f \colon M \to \mathbb{R}$ be continuous on M. Then f is bounded on M.

Definition. We say that a function f of n variables has a **limit** at a point $\vec{a} \in \mathbb{R}^n$, which is equal to $A \in \mathbb{R}^*$ if and only if

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0 \; \exists \delta \in \mathbb{R}, \delta > 0 \; \forall \vec{x} \in B(\vec{a}, \delta) \setminus \{\vec{a}\} \colon f(\vec{x}) \in B(A, \varepsilon).$$

We denote it $\lim_{\vec{x}\to\vec{a}} f(\vec{x}) = A$.

Remarks. 1. If the function f should have a limit at a point $\vec{a}$ according to our definition, it must exist $\delta_0 > 0$ such that f is defined at each point of the set $B(\vec{a}, \delta_0) \setminus \{\vec{a}\}$. We could extend this definition by introducing a notion of the limit of the function at a point in a set. However, we do not need this extension in the following reading.

2. Every function has at a fixed point at most one limit.

3. Notice that $\lim_{\vec{x}\to\vec{a}} f(\vec{x}) = f(\vec{a})$ if and only if f is continuous at $\vec{a}$.

For limits of multivariate functions holds similar theorems to theorems about limits for for functions of one real variable (e.g. theorem about limit arithmetic or sandwich theorem). We will formulate explicitly one variant of the theorem about the limit of the composite function.

Theorem 21. Let $r, s \in \mathbb{N}$, $\vec{a} \in M \subset \mathbb{R}^s$, $L \subset \mathbb{R}^r$, $\varphi_1, \ldots, \varphi_r$ be functions defined on M satisfying $\lim_{\vec{x}\to\vec{a}} \varphi_j(\vec{x}) = b_j$, $j = 1, \ldots, r$, a $\vec{b} = [b_1, \ldots, b_r] \in L$. Let $f\colon L \to \mathbb{R}$ is continuous at a point $\vec{b}$. Define composite function $F\colon M \to \mathbb{R}$ by

$$F(\vec{x}) = f\big(\varphi_1(\vec{x}), \varphi_2(\vec{x}), \ldots, \varphi_r(\vec{x})\big), \quad \vec{x} \in M.$$

Then $\lim_{\vec{x}\to\vec{a}} F(\vec{x}) = f(\vec{b})$.

Example 22. Determine the domain of the function $f(x, y) = \sqrt{\log(x - y)}$; examine continuity of the function and draw some contour lines (i.e. sets $f_{-1}(\{c\})$, $c \in \mathbb{R}$).

Solution. The domain of the function f is a set

$$D_f = \{[x, y] \in \mathbb{R}^2;\ \log(x - y) \geq 0\} = \{[x, y] \in \mathbb{R}^2;\ x - y \geq 1\}.$$

Since the coordinates projections π^1 and π^2 are continuous on the whole $\mathbb{R}^2$ (see Example 14), is the function $g(x, y) = \pi^1(x, y) - \pi^2(x, y) = x - y$ continuous on the whole $\mathbb{R}^2$ (and thus also on $D_f \subset \mathbb{R}^2$). The function $\log$ is continuous on $(0, +\infty)$, and therefore the function $\log(x - y)$ is continuous on D_f (composition of the continuous functions). The function $u \mapsto \sqrt{u}$ is continuous on $[0, +\infty)$, and thus the function $f(x, y) = \sqrt{\log(x - y)}$ is also continuous on D_f.

Since the function f is non-negative, $f_{-1}(\{c\}) = \emptyset$ holds for each $c < 0$. Set c equal the numbers 0, $\sqrt{\log 2}$ and $\sqrt{\log 3}$ one by one:

$$f_{-1}(\{0\}) = \{[x, y] \in \mathbb{R}^2;\ x - y = 1\},$$
$$f_{-1}\big(\{\sqrt{\log 2}\}\big) = \{[x, y] \in \mathbb{R}^2;\ x - y = 2\},$$
$$f_{-1}\big(\{\sqrt{\log 3}\}\big) = \{[x, y] \in \mathbb{R}^2;\ x - y = 3\}.$$

Draw these sets in $\mathbb{R}^2$. ♣

Example 23. Examine continuity of the function defined by

$$f(x,y) = \begin{cases} \frac{2xy}{x^2+y^2} & \text{pro } [x,y] \neq [0,0], \\ 0 & \text{pro } [x,y] = [0,0] \end{cases}$$

in $\mathbb{R}^2$ and draw some of its contour lines.

Solution. Continuity at points of the open set $\mathbb{R}^2 \setminus \{[0,0]\}$ can be shown similarly to the way in the previous example.

From the Heine theorem follows very easily that at the point $[0,0]$ the function f is not continuous: take an arbitrary sequence $\{a_n\} \subset \mathbb{R}$ such that, $\lim a_n = 0$ and $a_n \neq 0$. The sequence $\{[a_n, a_n]\}$ converges to the point $[0,0]$ in $\mathbb{R}^2$, but $\lim f(a_n, a_n) = 1 \neq 0 = f(0,0)$.

If we realize that $(x \pm y)^2 \geq 0$ holds for all pairs of the real numbers x and y, we get immediately, that for all points $[x,y] \in \mathbb{R}^2 \setminus \{[0,0]\}$ is $-1 \leq \frac{2xy}{x^2+y^2} \leq 1$. The function f is thus bounded and we see that for each $c < -1$ and for each $c > 1$ also $f_{-1}(\{c\}) = \emptyset$ holds.

Try to formulate some of the contour lines:

$$f_{-1}(\{-1\}) = \{[x,y] \in \mathbb{R}^2;\ y = -x\} \setminus \{[0,0]\},$$

$$f_{-1}(\{-1/\sqrt{2}\}) =$$

$$= \{[x,y] \in \mathbb{R}^2;\ (y + \sqrt{2}x - x)(y + \sqrt{2}x + x) = 0\} \setminus \{[0,0]\} =$$

$$= \{[x,y] \in \mathbb{R}^2;\ y = (1-\sqrt{2})x \vee y = (-1-\sqrt{2})x\} \setminus \{[0,0]\},$$

$$f_{-1}(\{0\}) = \{[x,y] \in \mathbb{R}^2;\ x = 0\} \cup \{[x,y] \in \mathbb{R}^2;\ y = 0\},$$

$$f_{-1}(\{1/\sqrt{2}\}) = \{[x,y] \in \mathbb{R}^2;\ y = (\sqrt{2}+1)x \vee y = (\sqrt{2}-1)x\} \setminus \{[0,0]\},$$

$$f_{-1}(\{1\}) = \{[x,y] \in \mathbb{R}^2;\ y = x\} \setminus \{[0,0]\}.$$

♣

Example 24. Examine continuity of the function defined on $\mathbb{R}^2$ by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & \text{for } [x,y] \neq [0,0], \\ 0 & \text{for } [x,y] = [0,0], \end{cases}$$

and determine if the function has a maximum and minimum on $\mathbb{R}^2$.

Solution. Continuity of the function is obvious at all points except the origin. Examine continuity of the function at th point $[0,0]$. We will estimate the subtraction $|f(x,y) - f(0,0)| = |f(x,y)|$. For each point $[x,y] \in \mathbb{R}^2 \setminus \{[0,0]\}$ it is

$$|f(x,y)| \leq \frac{(x^2+y^2)\,|y|}{x^2+y^2} = |y|\,.$$

Let $\varepsilon > 0$. For each $[x, y] \in B([0, 0], \varepsilon)$ it is

$$|f(x, y)| \leq |y| \leq \sqrt{x^2 + y^2} < \varepsilon.$$

Given any positive number $\varepsilon > 0$ we cab find $\delta > 0$ (for example $\delta = \varepsilon$) such that for $[x, y] \in B([0, 0], \delta)$ is $|f(x, y) - f(0, 0)| < \varepsilon$, that means that the function f is continuous at the point $[0, 0]$.

The function f is not bounded above on $\mathbb{R}^2$, because for each $c > 0$ there exists a point $[x, x]$ such that $f(x, x) = x/2 > c$. Similarly, we can prove, that the function f is not bounded below. Since f is bounded on $\mathbb{R}^2$ neither above nor below, it has neither maximum nor minimum on that set. ♣

Example 25. Determine the domain of the function $f(x, y) = \sqrt{\frac{1}{x^2} - \frac{y^2}{x^2} - 1}$. Examine continuity of the function on D_f and maximum and minimum of the function f and draw some of its contour lines.

Solution. The domain is $D_f = \{[x, y] \in \mathbb{R}^2;\ x \neq 0, 1 - y^2 - x^2 \geq 0\}$. It is then a circle with radius 1 with points, where x-coordinate is equal to zero (part of the y-axis), being removed. Continuity on the domain can be proved similarly using composition of continuous functions.

Determine some contour lines:

$$f_{-1}(\{0\}) = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 = 1\} \setminus \{[0, 1], [0, -1]\},$$
$$f_{-1}(\{1\}) = \{[x, y] \in \mathbb{R}^2;\ 2x^2 + y^2 = 1\} \setminus \{[0, 1], [0, -1]\},$$
$$f_{-1}(\{2\}) = \{[x, y] \in \mathbb{R}^2;\ 5x^2 + y^2 = 1\} \setminus \{[0, 1], [0, -1]\}.$$

The function attains on D_f the smallest value 0 at each point of the contour line $f_{-1}(\{0\})$. It does not have a maximum, because is not bounded above. If we are getting closer to the point $[0, 0]$ along the x-axis, we get $\lim_{x \to 0} f(x, 0) = \sqrt{\frac{1}{x^2} - 1} = +\infty$. ♣

Example 26. Determine the distance from the point $[-5, -1]$ to the sets

$$M_1 = \{[x, y] \in \mathbb{R}^2;\ y = x^2\} \quad \text{a} \quad M_2 = \{[x, y] \in \mathbb{R}^2;\ y > x^2\}.$$

Solution. We know the notions of the distance from a point to a line and the distance from a point to a plane from the high school: we find the closest point on the line (the plane, respectively) to a given point and the searched distance is the distance between these two points. If we want to define the distance from a point $\vec{a} \in \mathbb{R}^n$ to a general set $M \subset \mathbb{R}^n$, we find a difficulty that the closest point does not have to be included in M. Thus we define the distance from a point to the set in a following way. Let $\vec{a} \in \mathbb{R}^n$ and $M \subset \mathbb{R}^n$, $M \neq \emptyset$. Then we call the number

$$\rho(\vec{a}, M) = \inf \{\rho(\vec{a}, \vec{x});\ \vec{x} \in M\}$$

the **distance from the point $\vec{a}$ to the set M**.

According to this definition, in our example then follows

$$\rho([-5,-1],M_1)=\inf\big\{\rho([-5,-1],[x,x^2]);\ x\in\mathbb{R}\big\}=$$
$$=\inf\big\{\sqrt{(x+5)^2+(x^2+1)^2};\ x\in\mathbb{R}\big\}.$$

We can find that function $f(x)=\sqrt{(x+5)^2+(x^2+1)^2}$ on $\mathbb{R}$ has a minimum at the point $x=-1$. Thus $\rho([-5,-1],M_1)=f(-1)=2\sqrt{5}$. In this case the set M_1 contains the point $[-1,1]$, which is closest to the $[-5,-1]$. Prove on your own, that $\rho([-5,-1],M_2)=2\sqrt{5}$, and that the set M_2 does not contain the closest point of the point $[-5,-1]$. ♣

1.3. Partial derivative and tangent hyperplane

In this section we will show, how to generalize the notion the derivation of a function of one variable for the multivariate functions.

Definition. Let f be a function of n variables, $j\in\{1,\dots,n\}$ a $\vec{a}\in\mathbb{R}^n$. Then we call the number

$$\frac{\partial f}{\partial x_j}(\vec{a})=\lim_{t\to 0}\frac{f(\vec{a}+t\vec{e^j})-f(\vec{a})}{t}$$
$$\left(=\lim_{t\to 0}\frac{f(a_1,\dots,a_{j-1},a_j+t,a_{j+1},\dots,a_n)-f(a_1,\dots,a_n)}{t}\right)$$

the **(first order) partial derivative of the function f with respect to the j-th variable at the point a $\vec{a}$** (provided that the limit exists).

Remarks. 1. If $\frac{\partial f}{\partial x_j}(\vec{a})$ should exist, then there must exists $\delta>0$ such that line segment $\{\vec{a}+t\vec{e^j};\ |t|\le\delta\}$ is a subset of D_f.

2. Setting

$$g(y)=f(a_1,a_2,\dots,a_{j-1},y,a_{j+1},\dots,a_n),$$

then $g'(a_j)$ exists, if and only if there exists $\frac{\partial f}{\partial x_j}(\vec{a})$. If both derivatives exist then they are equal. On the following figure we can see the geometric meaning of the function g.

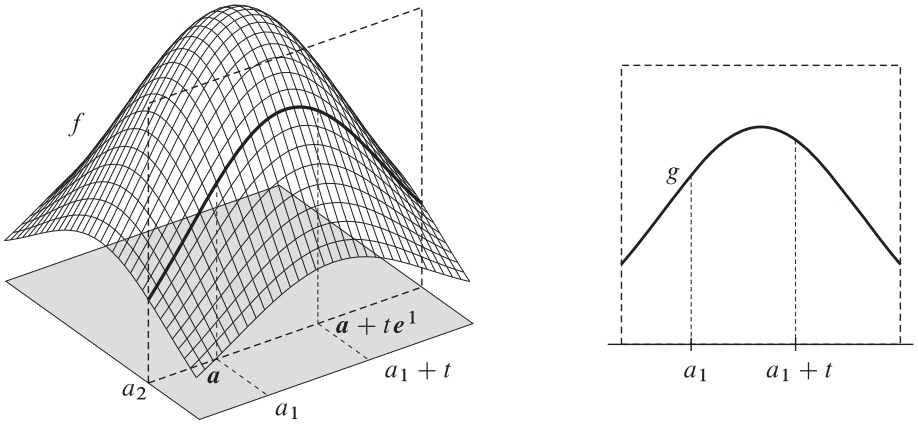


FIGURE 1.

This remark gives us instruction, how to calculate the partial derivatives of the multivariate functions with the use of the derivatives of the function of one variable.

The value of the notion of a partial derivative will be already demonstrated in the following theorem.

Theorem 27 (necessary condition for a local extrem)**.** Let $G \subset \mathbb{R}^n$ be open, $\vec{a} \in G$ and a function $f\colon G \to \mathbb{R}$ has a local extrem (on G) at the point $\vec{a}$. Then for each $j \in \{1, \dots, n\}$ holds: the partial derivative $\frac{\partial f}{\partial x_j}(\vec{a})$ either does not exist, or is equal to zero.

Proof. Choose $j \in \{1, \dots, n\}$ and set $g(t) = f(\vec{a} + t\vec{e^j})$. The function g is defined on some neighbourhood of 0 and at the point 0 it has a local extrem. According to the Theorem **??** we obtain that the derivative $g'(0)$ either does not exist, or is equal to zero, and because $g'(0) = \frac{\partial f}{\partial x_j}(\vec{a})$holds (provided that at least one limit exists), the proof is finished. ∎

Partial derivatives are a very useful tool in examining the properties of the multivariate functions. However, there is a disadvantage that every partial derivative at some point describe behaviour of the derived function only in one particular direction.

If we consider only functions, which have continuous all partial derivatives on the open set $G \subset \mathbb{R}^n$, then partial derivatives give us more complex knowledge about behaviour of the original function (see e.g. Theorem 30 and Theorem 31).

Definition. Let $G \subset \mathbb{R}^n$ be nonempty and open. Let the function $f\colon G \to \mathbb{R}$ has continuous all partial derivatives at each point of the set G (that is, the function $\vec{x} \mapsto \frac{\partial f}{\partial x_j}(\vec{x})$, $j = 1, \dots, n$, are continuous at each point of G). Then we say that

the function f is of **class $\mathcal{C}^1$ on G**. We denote the set of all such functions f by $\mathcal{C}^1(G)$.

Remark. Let $G \subset \mathbb{R}^n$ is nonempty open set and $f, g \in \mathcal{C}^1(G)$. Then th functions $f+g, f-g, fg$ are also of the class $\mathcal{C}^1(G)$. Provided that g does not attain zero value in any point of the set G, then also $f/g \in \mathcal{C}^1(G)$.

Remark. An important task is finding extremes of the function f on K. We are often in situation that $K \subset \mathbb{R}^n$ is closed nonempty bounded set, f is continuous function on K and is of the class $\mathcal{C}^1$ on the interior of the set K.

According to the characterization of the compact sets in $\mathbb{R}^n$ (Theorem 17), we get that the set K is compact, and hence f attains its maximum and minimum values on K (Theorem 19). Each point of the extrem is also a local extrem point on K and lies either on the border, or in the interior of the set K. If there is an extrem at a point $\vec{x} \in \operatorname{Int} K$ then applies

$$\frac{\partial f}{\partial x_j}(\vec{x}) = 0,\ j = 1, \ldots, n. \tag{7}$$

The function f could attain a extrem value only at points which satisfy the condition (7) or lies on the border of K. We introduce the methods, which enables to find points on the border suspicious to be an extrem, in the following example, and particularly in section 1.5, where we show one general example. Since the necessary condition for local extremes are often satisfied for only a finite number of points of the set K, then it is sufficient to calculate values of the function in these points, compare them and thus find extremes.

Example 28. Find extremes of the function $f(x,y) = 3x^2 + 4y^3$ on the set $M = \{[x,y] \in \mathbb{R}^2;\ x^2 + y^2 \leq 1\}$.

Solution. The function f is continuous on the whole $\mathbb{R}^2$ and thus is continuous on the set M. The set M is bounded and closed (see the Theorem 16), and therefore it is compact. The function f has a maximum and minimum on it. The suspicious points (i.e. the points, at which only could be an extrem) are the points of the border M and – since $f \in \mathcal{C}^1(\operatorname{Int} M)$ – points inside M, satisfying the condition (7).

Find all suspicious points inside M at first. Calculate both first order partial derivative:

$$\frac{\partial f}{\partial x}(x,y) = 6x, \qquad \frac{\partial f}{\partial y}(x,y) = 12y^2.$$

To find points, where both partial derivative equal 0, we have to solve the linear system:

$$\begin{aligned} 6x &= 0, \\ 12y^2 &= 0. \end{aligned}$$

The function f has both partial derivative equals 0 only at the point $[0,0]$, which lies in the set $\operatorname{Int} M$.

Now determine the points on the border. Define a supporting partial function φ, which maps an interval $[0,2\pi]$ on $H(M)$:

$$\varphi(t) = [\cos t, \sin t], \quad t \in [0, 2\pi].$$

Define a function $g\colon [0,2\pi] \to \mathbb{R}$ in a following way: $g(t) = f\big(\varphi(t)\big)$, $t \in [0,2\pi]$. Thus $g(t) = 3\cos^2 t + 4\sin^3 t$ holds for the function g. Find a maximum of the function g on the interval $[0,2\pi]$: $g'(t) = -6\cos t \sin t + 12\sin^2 t \cos t$ holds. Inside the interval $[0,2\pi]$ is the derivation of the function g equal to zero at the points $\pi/6$, $\pi/2$, $5\pi/6$, π, $3\pi/2$. The function g is continuous on bounded and closed interval $[0,2\pi]$, and hence it has a maximum and minimum on the interval. Extrém tedy může mít pouze v bodech, kde $g'(t) = 0$, a v krajních bodech intervalu $[0,2\pi]$. It is easily determined, that function g has a maximum on the interval at the point $\pi/2$ and a minimum at the point $3\pi/2$.

Now return to the function f. From the previous paragraph it follows, That the extrem could by only at points $[0,0]$, $[0,1]$, $[0,-1]$. For the function values of f at these points $f(0,0) = 0$, $f(0,1) = 4$, $f(0,-1) = -4$ holds. Hence, the function f has a maximum on M at the point $[0,1]$ and has there the function value equal to 4 and f has a minimum on M at the point $[0,-1]$ and has the function value equal to -4. ♣

Convention. Let $a, b \in \mathbb{R}$. Then we denote a closed interval with endpoints a a b by the symbol $[a,b]$, also in the case, where $a > b$.

Theorem 29 (weak version of the Lagrange theorem)**.** Let $n \in \mathbb{N}$, $I_1, \ldots, I_n \subset \mathbb{R}$ be open intervals, $I = I_1 \times \cdots \times I_n$, $f \in \mathcal{C}^1(I)$, $\vec{a}, \vec{b} \in I$. Then we can find points $\vec{\xi^i} \in I$, $i = 1, \ldots, n$, satisfying $\xi^i_j \in [a_j, b_j]$ for all $i, j \in \{1, \ldots, n\}$ such that,

$$f(\vec{b}) - f(\vec{a}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{\xi^i})(b_i - a_i).$$

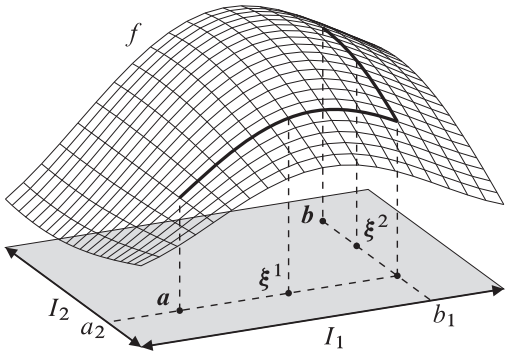


FIGURE 2.

Proof. Firstly, proof the assertion for $n = 1$. If $a = b$, choose $\xi^1 = a$. If $a < b$, is the function f continuous on interval $[a, b]$ (since it is differentiable on I) and has a derivation at each point of the (a, b), thus the Lagrange theorem (Theorem **??**) can be used, from it we obtain that exists $\xi^1 \in (a, b)$ such that

$$f(b) - f(a) = f'(\xi^1)(b - a) = \frac{\partial f}{\partial x_1}(\xi^1)(b - a).$$

If $a > b$, we can get similarly the existence of the number $\xi^1 \in (b, a)$ satisying $f(a) - f(b) = f'(\xi^1)(a - b)$, thas $f(b) - f(a) = \frac{\partial f}{\partial x_1}(\xi^1)(b - a)$ also holds.

Then we follow by proving the theorem for $n = 2$. Define functions $g_1 \colon I_1 \to \mathbb{R}$ and $g_2 \colon I_2 \to \mathbb{R}$ by $g_1(t) = f(t, a_2)$ and $g_2(t) = f(b_1, t)$. Then

$$\begin{aligned} f(\vec{b}) - f(\vec{a}) &= f(b_1, b_2) - f(b_1, a_2) + f(b_1, a_2) - f(a_1, a_2) = \\ &= g_2(b_2) - g_2(a_2) + g_1(b_1) - g_1(a_1). \end{aligned} \tag{8}$$

The function g_1 is of the class $\mathcal{C}^1$ on I_1 and the function g_2 is of the class $\mathcal{C}^1$ on I_2. According to the first part of the proof there exist numbers $\xi_1^1 \in [a_1, b_1]$ and $\xi_2^2 \in [a_2, b_2]$ such that,

$$\begin{aligned} g_1(b_1) - g_1(a_1) &= g_1'(\xi_1^1)(b_1 - a_1) = \frac{\partial f}{\partial x_1}(\xi_1^1, a_2)(b_1 - a_1), \\ g_2(b_2) - g_2(a_2) &= g_2'(\xi_2^2)(b_2 - a_2) = \frac{\partial f}{\partial x_2}(b_1, \xi_2^2)(b_2 - a_2). \end{aligned}$$

Setting $\vec{\xi^1} = [\xi_1^1, a_2]$ a $\vec{\xi^2} = [b_1, \xi_2^2]$, we get

$$f(\vec{b}) - f(\vec{a}) = \frac{\partial f}{\partial x_1}(\vec{\xi^1})(b_1 - a_1) + \frac{\partial f}{\partial x_2}(\vec{\xi^2})(b_2 - a_2).$$

from the (8).

Finally for $n > 2$ we have

$$f(\vec{b}) - f(\vec{a}) = \sum_{i=1}^{n} \big(f(b_1, \ldots, b_{i-1}, b_i, a_{i+1}, \ldots, a_n) - \\ - f(b_1, \ldots, b_{i-1}, a_i, a_{i+1}, \ldots, a_n) \big)$$

and the proof can be completed similarly to the one in the previous step. ∎

Definition. Let $G \subset \mathbb{R}^n$ be an open set, $\vec{a} \in G$ and $f \in \mathcal{C}^1(G)$. Then we call the graph of the function $T\colon \mathbb{R}^n \to \mathbb{R}$ defined by the formula

$$T(\vec{x}) = f(\vec{a}) + \frac{\partial f}{\partial x_1}(\vec{a})(x_1 - a_1) + \cdots + \frac{\partial f}{\partial x_n}(\vec{a})(x_n - a_n) \tag{9}$$

a **tangent hyperplane to the graph** of the function f at the point $[\vec{a}, f(\vec{a})]$.

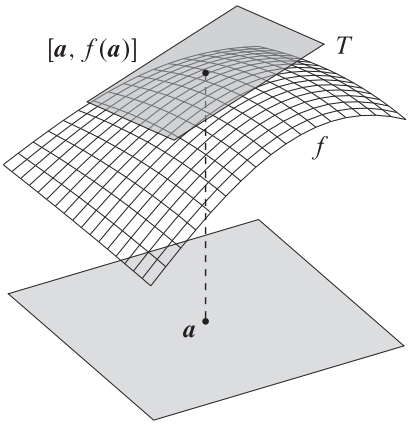


FIGURE 3.

Observe that the point $[\vec{a}, f(\vec{a})]$ is an element of the tangent hyperplane. Notice that in the case $n = 1$ and $n = 2$ respectively, we say tangent (see the chapter **??**) and **tangent plane**, respectively, instead of tangent hyperplane.

The following theorem justify the use of adjective "tangent". It says that, tha error we make by replacing the value $f(\vec{x})$ by the value $T(\vec{x})$, approach 0 faster than $\rho(\vec{x}, \vec{a})$ for $\vec{x}$ approaching $\vec{a}$.

Theorem 30 (about tangent hyperplane)**.** Let $G \subset \mathbb{R}^n$ be an open set, $\vec{a} \in G$, $f \in \mathcal{C}^1(G)$ and T be a function, whose graph is a tnagent hyperplane to the graph of the function f at the point $[\vec{a}, f(\vec{a})]$. Then

$$\lim_{\vec{x} \to \vec{a}} \frac{f(\vec{x}) - T(\vec{x})}{\rho(\vec{x}, \vec{a})} = 0.$$

Proof. Let $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$ by given. Since the function f has all partial derivatives continuous at the point $\vec{a}$, we can find $\Delta > 0$ such that $B(\vec{a}, \Delta) \subset G$ and for each $i \in \{1, \ldots, n\}$ and each $\vec{x} \in B(\vec{a}, \Delta)$ is $\left|\frac{\partial f}{\partial x_i}(\vec{x}) - \frac{\partial f}{\partial x_i}(\vec{a})\right| < \frac{\varepsilon}{n}$. Setting $\delta = \Delta/\sqrt{n}$ and $I = (a_1-\delta, a_1+\delta)\times\cdots\times(a_n-\delta, a_n+\delta)$, we obtain $I \subset B(\vec{a}, \Delta)$, what we can calculate easily. For each $i \in \{1, \ldots, n\}$ and each $\vec{x} \in I$ is thus

$$\left|\frac{\partial f}{\partial x_i}(\vec{x}) - \frac{\partial f}{\partial x_i}(\vec{a})\right| < \frac{\varepsilon}{n}. \tag{10}$$

Let $\vec{y} \in B(\vec{a}, \delta)$ by given. Then $\vec{y} \in I$ and thus according to the Theorem 29 there exist points $\vec{\xi^1}, \ldots, \vec{\xi^n} \in I$ such that

$$f(\vec{y}) - f(\vec{a}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{\xi^i})(y_i - a_i).$$

This yields

$$\begin{aligned}
\left|\frac{f(\vec{y}) - T(\vec{y})}{\rho(\vec{y}, \vec{a})}\right| &= \frac{\left|f(\vec{y}) - f(\vec{a}) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a})(y_i - a_i)\right|}{\rho(\vec{y}, \vec{a})} = \\
&= \frac{\left|\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}(\vec{\xi^i}) - \frac{\partial f}{\partial x_i}(\vec{a})\right)(y_i - a_i)\right|}{\rho(\vec{y}, \vec{a})} \leq \\
&\leq \frac{\sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(\vec{\xi^i}) - \frac{\partial f}{\partial x_i}(\vec{a})\right| |y_i - a_i|}{\rho(\vec{y}, \vec{a})} < \\
&< \frac{\varepsilon}{n} \cdot \frac{\sum_{i=1}^{n} |y_i - a_i|}{\rho(\vec{y}, \vec{a})} \leq \frac{\varepsilon}{n} \cdot \frac{\sum_{i=1}^{n} \rho(\vec{y}, \vec{a})}{\rho(\vec{y}, \vec{a})} = \varepsilon.
\end{aligned}$$

■

Theorem 31. Let $G \subset \mathbb{R}^n$ be open and $f \in \mathcal{C}^1(G)$. Then f is continuous on G.

Proof. Let $\vec{a} \in G$ and T be a function in a form (9). According to the Theorem 30 then follows

$$\lim_{\vec{x}\to\vec{a}} f(\vec{x}) = \lim_{\vec{x}\to\vec{a}} \left(\frac{f(\vec{x}) - T(\vec{x})}{\rho(\vec{x}, \vec{a})} \cdot \rho(\vec{x}, \vec{a}) + T(\vec{x})\right) = 0 \cdot 0 + f(\vec{a}) = f(\vec{a}).$$

According to the remark on the page 13, the function f is thus continuous at the point $\vec{a}$. ■

Now we will introduce an analogy of the theorem about the derivative of the composite function.

Theorem 32 (derivative of the composite function)**.** Let $r, s \in \mathbb{N}$ and $G \subset \mathbb{R}^s$, $H \subset \mathbb{R}^r$ be open sets. Let $\varphi_1, \dots, \varphi_r \in \mathcal{C}^1(G)$, $f \in \mathcal{C}^1(H)$ and for each $\vec{x} \in G$ is the point $[\varphi_1(\vec{x}), \dots, \varphi_r(\vec{x})] \in H$. Then the composite function $F\colon G \to \mathbb{R}$ defined by a formula

$$F(\vec{x}) = f\big(\varphi_1(\vec{x}), \varphi_2(\vec{x}), \dots, \varphi_r(\vec{x})\big), \quad \vec{x} \in G,$$

is of the class $\mathcal{C}^1$ on G. Let $\vec{a} \in G$ and $\vec{b} = [\varphi_1(\vec{a}), \dots, \varphi_r(\vec{a})]$. Then

$$\frac{\partial F}{\partial x_j}(\vec{a}) = \sum_{i=1}^{r} \frac{\partial f}{\partial y_i}(\vec{b}) \frac{\partial \varphi_i}{\partial x_j}(\vec{a}) \tag{11}$$

holds for $j \in \{1, \dots, s\}$.

Remark. The symbol $\frac{\partial f}{\partial y_i}(\vec{b})$ in the formula (11) denotes the partial derivative of the function f with respect to the i-th variable at the point $\vec{b}$.

Proof. According to the remark on the page 16 we can assume without loss of generality that $s = 1$. We will calculate the derivative of the function F at the point $a \in \mathbb{R}$, that is the limit $\lim\limits_{x \to a} \frac{F(x)-F(a)}{x-a}$.

The set H is open, and hence there exists $\Delta > 0$ such that $B(\vec{b}, \Delta) \subset H$. Now we can find open intervals $I_1, \dots, I_r \subset \mathbb{R}$ such that

$$\vec{b} \in I = I_1 \times \cdots \times I_r \subset H.$$

It can be easily shiwn by calculation, that the choice $I_i = \big(b_i - \Delta/\sqrt{r}, b_i + \Delta/\sqrt{r}\big)$, $i = 1, \dots, r$ satisfy the condition . Since the set G be open and the functions $\varphi_1, \dots, \varphi_r$ are continuous, there exists $\delta \in \mathbb{R}$, $\delta > 0$, such that $(a - \delta, a + \delta) \subset G$ and for each $x \in (a - \delta, a + \delta)$ and $i \in \{1, \dots, r\}$ is $\varphi_i(x) \in I_i$. From the Theorem 29 thus follows, that for each $x \in (a - \delta, a + \delta)$ exist points $\vec{\xi^i}(x) \in I$, $i = 1, \dots, r$, satisfying $\xi^i(x)_j \in [b_j, \varphi_j(x)]$ for all $i, j \in \{1, \dots, r\}$ and

$$f\big(\varphi_1(x), \dots, \varphi_r(x)\big) - f(b_1, \dots, b_r) = \sum_{i=1}^{r} \frac{\partial f}{\partial y_i}\big(\vec{\xi^i}(x)\big)(\varphi_i(x) - b_i).$$

This way we defined the functions $x \mapsto \xi^i(x)_j$, $i = 1, \dots, r$, $j = 1, \dots, r$ on the interval $(a - \delta, a + \delta)$.

This yields

$$\begin{aligned} \frac{F(x) - F(a)}{x - a} &= \frac{f\big(\varphi_1(x), \dots, \varphi_r(x)\big) - f(b_1, \dots, b_r)}{x - a} = \\ &= \sum_{i=1}^{r} \frac{\partial f}{\partial y_i}\big(\vec{\xi^i}(x)\big) \frac{\varphi_i(x) - \varphi_i(a)}{x - a}. \end{aligned}$$

From the continuity of the functions φ_j at the point a (according to the Theorem **??**,eventually the Theorem 31), we obtain $\lim_{x\to a}\varphi_j(x) = b_j$. According to the sandwich theorem (Theorem **??**(iii)) is thus $\lim_{x\to a}\xi^i(x)_j = b_j$ for all $i,j \in \{1,\dots,r\}$. The functions $\frac{\partial f}{\partial y_i}$ are continuous at $\vec{b}$, hence due to the theorem about the limit of the composite function (Theorem 21) we get

$$F'(a) = \lim_{x\to a}\frac{F(x)-F(a)}{x-a} = \sum_{i=1}^{r}\frac{\partial f}{\partial y_i}(\vec{b})\cdot\varphi_i'(a).$$

■

Definition. Let $G \subset \mathbb{R}^n$ be an open set, $\vec{a} \in G$ a $f \in \mathcal{C}^1(G)$. We call a vector

$$\nabla f(\vec{a}) = \left[\frac{\partial f}{\partial x_1}(\vec{a}), \frac{\partial f}{\partial x_2}(\vec{a}), \dots, \frac{\partial f}{\partial x_n}(\vec{a})\right].$$

a **gradient of the function f at the point $\vec{a}$**

Remark. A gradient of the function sometimes helps us to know behaviour of the function better, because it determine the direction of the biggest growth at the point in the following sense. LEt $G \subset \mathbb{R}^n$ be an open set, $\vec{a} \in G$ and $f \in \mathcal{C}^1(G)$. If $\nabla f(\vec{a}) \neq \vec{o}$, $\vec{v} \in \mathbb{R}^n$, $\vec{v} \neq \nabla f(\vec{a})$ and $\rho(\vec{v},\vec{o}) = \rho\big(\nabla f(\vec{a}),\vec{o}\big)$, then it can be shown, that exists $\delta > 0$ such that

$$\forall t \in \mathbb{R}, t \in (0,\delta)\colon f\big(\vec{a} + t\nabla f(\vec{a})\big) > f(\vec{a} + t\vec{v}).$$

On the first figure, there is a graph of some function f and on the second, there are gradients $\nabla f(\vec{a})$ drawn on yhe plane for some values of $\vec{a} \in \mathbb{R}^2$. Notice the mutual relationship of both figures.

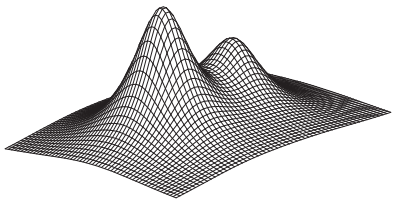


FIGURE 4. The graph of the function f

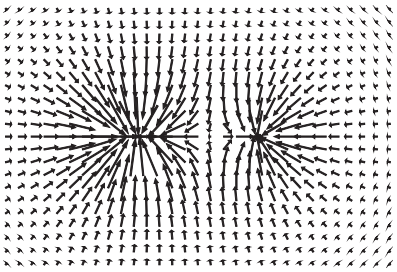


FIGURE 5. The gradient array

Definition. We call the point $\vec{a} \in \mathbb{R}^n$ satisfying $\nabla f(a) = \vec{o}$ a **stationary** (sometimes also **critical**) point of the function f.

We could define higher order partial derivative similarly to the definition of the higherof the function of one real variable.

Definition. Let $G \subset \mathbb{R}^n$ be a nonempty open set, $i, j \in \{1, \dots, n\}$, the function $f\colon G \to \mathbb{R}$ has a real i-th partial derivative at each point of G and $\vec{a} \in G$. We denote a partial derivative of the function $\vec{x} \mapsto \frac{\partial f}{\partial x_i}(\vec{x})$ with respect to x_j at the point $\vec{a}$ by

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a}) = \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(\vec{a})$$

and we call it a **second order partial derivative** of the function f. If $i = j$, then we use a notation $\frac{\partial^2 f}{\partial x_i^2}(\vec{a})$. We define a higher order partial derivatives analogically.

Generally, it matters if we derive firstly with respect to i-th and then with respect to j-th variable or conversely. However, the following theorem holds. We omit its (rather difficultier) proof.

Theorem 33. Let $i, j \in \mathbb{N}$, $i \leq n$, $j \leq n$, and the function f has both derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$, $\frac{\partial^2 f}{\partial x_j \partial x_i}$ on the neighbourhood of the point $\vec{a} \in \mathbb{R}^n$. These derivative are continuous at the point $\vec{a}$. Then

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{a}).$$

We will end this section by one more definition.

Definition. Let $G \subset \mathbb{R}^n$ be an open set and $k \in \mathbb{N}$. We say that a function f is **of the class $\mathcal{C}^k$ on G**, provided that f has all partial derivative up to order k and they are continuous on the set G. We denote the set of all functions of the class $\mathcal{C}^k$ on the set G by $\mathcal{C}^k(G)$.

We say that a function f is **of the class $\mathcal{C}^\infty$ on G**, provided that f has all partial derivative of all orders and they are continuous on the set G. We denote the set of all functions of the class $\mathcal{C}^\infty$ on G by $\mathcal{C}^\infty(G)$.

If we say that the function f is of the class $\mathcal{C}^k$ (without mentioning the set), it means that the function is defined on some nonempty open set G and is of the class $\mathcal{C}^k$ on G. We introduce similar convention also for $\mathcal{C}^\infty$.

Example 34. Determine the domain of the function $f(x, y) = \sqrt{|xy|}$. Calculate the partial derivatives at every point, where they exist. Find a tangent plane to the graph of the function f at the point $[1, -2, \sqrt{2}]$.

Solution. The domain of the function f is $\mathbb{R}^2$ and the function f is continuous on $\mathbb{R}^2$. We can rewrite $f(x, y) = \sqrt{|x|} \cdot \sqrt{|y|}$ and then we can easily calculate:

$$\frac{\partial f}{\partial x}(x, y) = \sqrt{|y|} \cdot \frac{\operatorname{sgn} x}{2\sqrt{|x|}} \quad \text{pro } [x, y] \in \mathbb{R}^2,\ x \neq 0,$$

$$\frac{\partial f}{\partial y}(x, y) = \sqrt{|x|} \cdot \frac{\operatorname{sgn} y}{2\sqrt{|y|}} \quad \text{pro } [x, y] \in \mathbb{R}^2,\ y \neq 0.$$

At the points, where previous formulas does not hold, calculate the partial derivatives of the function f from the definition. Firstly, calculate both partial derivative at the point $[0,0]$:

$$\frac{\partial f}{\partial x}(0,0) = \lim_{x\to 0} \frac{f(x,0)-f(0,0)}{x-0} = \lim_{x\to 0} 0 = 0,$$
$$\frac{\partial f}{\partial y}(0,0) = \lim_{y\to 0} \frac{f(0,y)-f(0,0)}{y-0} = \lim_{y\to 0} 0 = 0.$$

If $y_0 \neq 0$, then

$$\frac{\partial f}{\partial x}(0,y_0) = \lim_{x\to 0} \frac{f(x,y_0)-f(0,y_0)}{x-0} =$$
$$= \lim_{x\to 0} \frac{\sqrt{|xy_0|}}{x} = \lim_{x\to 0} \frac{\operatorname{sgn} x}{\sqrt{|x|}} \sqrt{|y_0|};$$

this limit does not exist, and thus $\frac{\partial f}{\partial x}(0,y_0)$ does not exist. For $x_0 \neq 0$ we can prove similarly, that $\frac{\partial f}{\partial y}(x_0,0)$ does not exist.

The function f has continuous first order partial derivative on the neigbourhood of the point $[1,-2]$. The tangent plane at the point $[1,-2,\sqrt{2}]$ is the graph of the function

$$T(x,y) = \sqrt{2} + \frac{1}{\sqrt{2}}(x-1) - \frac{1}{2\sqrt{2}}(y+2).$$

♣

Example 35. The function f is on $\mathbb{R}^2$ defined by

$$f(x,y) = \begin{cases} (x^2+y^2)\sin\left(\frac{1}{x^2+y^2}\right) & \text{pro } [x,y] \neq [0,0], \\ 0 & \text{pro } [x,y] = [0,0]. \end{cases}$$

Calculate partial derivatives at all points, where they exist.

Solution. The domain of the function f is the whole $\mathbb{R}^2$ and the function f is continuous on $\mathbb{R}^2$. For the points $[x,y] \neq [0,0]$ is

$$\frac{\partial f}{\partial x}(x,y) = 2x\sin\left(\frac{1}{x^2+y^2}\right) + (x^2+y^2)\cos\left(\frac{1}{x^2+y^2}\right)\frac{-2x}{(x^2+y^2)^2} =$$
$$= 2x\sin\left(\frac{1}{x^2+y^2}\right) - \frac{2x}{x^2+y^2}\cos\left(\frac{1}{x^2+y^2}\right),$$
$$\frac{\partial f}{\partial y}(x,y) = 2y\sin\left(\frac{1}{x^2+y^2}\right) - \frac{2y}{x^2+y^2}\cos\left(\frac{1}{x^2+y^2}\right).$$

Calculate partial derivatives at the point $[0,0]$ according to the definition:

$$\begin{aligned}\frac{\partial f}{\partial x}(0,0) &= \lim_{x\to 0}\frac{f(x,0)-f(0,0)}{x-0} = \lim_{x\to 0}\frac{x^2\sin\frac{1}{x^2}}{x} = \\ &= \lim_{x\to 0} x\sin\frac{1}{x^2} = 0,\\ \frac{\partial f}{\partial y}(0,0) &= 0.\end{aligned}$$

♣

Example 36. Determine the tangent plane at the point $[0,3,\sqrt{3}]$ to the torus defined by the equation $(x^2+y^2+z^2+12)^2-64(x^2+y^2)=0$.

Solution. By expressing z as a function of two variables we find out that the plane can be described as the union of two graphs of the following functions f and g:

$$\begin{aligned}f(x,y) &= \sqrt{8\sqrt{x^2+y^2}-x^2-y^2-12},\\ D_f &= \{[x,y]\in\mathbb{R}^2;\ 4\le x^2+y^2\le 36\},\\ g(x,y) &= -\sqrt{8\sqrt{x^2+y^2}-x^2-y^2-12},\\ D_g &= D_f.\end{aligned}$$

Since $f(0,3)=\sqrt{3}$, the point $[0,3,\sqrt{3}]$ is a point of the graph of the function f. The partial derivatives

$$\begin{aligned}\frac{\partial f}{\partial x}(x,y) &= x\frac{4-\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}\sqrt{8\sqrt{x^2+y^2}-x^2-y^2-12}},\\ \frac{\partial f}{\partial y}(x,y) &= y\frac{4-\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}\sqrt{8\sqrt{x^2+y^2}-x^2-y^2-12}}\end{aligned}$$

are continuous on the neigbourhood of the point $[0,3]$, hence at this point there exists a tangent plane and is described by the function T defined by

$$T(x,y)=\sqrt{3}+0\cdot(x-0)+\frac{1}{\sqrt{3}}(y-3)=\sqrt{3}+\frac{1}{\sqrt{3}}(y-3)=\frac{y}{\sqrt{3}}.$$

♣

In next examples we will examine extremes of the multivariate functions.

Example 37. Find local extrems of the function $f(x,y)=xy\log(x^2+y^2)$. Determine if the function f attains a maximum and a minimum values on D_f maxima a minima; if so, calculate them.

Solution. The domain of the function is the set $D_f = \mathbb{R}^2 \setminus \{[0,0]\}$. The function f is continuous on the whole D_f. It the point $[0,0]$ we have

$$\lim_{[x,y]\to[0,0]} xy\log(x^2+y^2) = 0.$$

For calculation of this limit we will use the following estimate $|xy| \le (x^2+y^2)/2$ and the result $\lim_{u\to 0+} u\log u = 0$ (Example **??**).

Since

$$\lim_{x\to+\infty} f(x,x) = \lim_{x\to+\infty}(x^2\log 2x^2) = +\infty,$$
$$\lim_{x\to+\infty} f(x,-x) = \lim_{x\to+\infty}(-x^2\log 2x^2) = -\infty,$$

we can see that the function f is bounded neither above nor below on D_f, and thus it can attaint neither maximum nor minimum value on D_f.

Find points, where the function f could attain a local extrem value. Calculate first order partial derivatives at first:

$$\frac{\partial f}{\partial x}(x,y) = y\log(x^2+y^2) + \frac{2x^2y}{x^2+y^2},$$
$$\frac{\partial f}{\partial y}(x,y) = x\log(x^2+y^2) + \frac{2xy^2}{x^2+y^2},$$

whenever $[x,y] \in D_f$. The partial derivative exists at all points of D_f, and thus we can find suspicious points of $D_f$by solving the system of linear equation.

$$y\left(\log(x^2+y^2) + 2\frac{x^2}{x^2+y^2}\right) = 0, \qquad x\left(\log(x^2+y^2) + 2\frac{y^2}{x^2+y^2}\right) = 0.$$

The first equation holds if and only if $y = 0$ or $\log(x^2+y^2) = -\frac{2x^2}{x^2+y^2}$, the second holds if and only if $x = 0$ or $\log(x^2+y^2) = -\frac{2y^2}{x^2+y^2}$. By testing all possibilities, we get the following suspicious points

$$[0,1],\ [0,-1],\ [1,0],\ [-1,0],$$
$$[1/\sqrt{2e}, 1/\sqrt{2e}],\ [1/\sqrt{2e}, -1/\sqrt{2e}],$$
$$[-1/\sqrt{2e}, 1/\sqrt{2e}],\ [-1/\sqrt{2e}, -1/\sqrt{2e}].$$

Now, draw the domain $D_f = \mathbb{R}^2\setminus\{[0,0]\}$ and the set K inside it $K = \{[x,y] \in D_f;\ f(x,y) = 0\}$. The set K jis the union of the coordinate axis without the origin and the circle with the centre $[0,0]$ and radius 1. The set K divides $\mathbb{R}^2$ at areas, kwhere the function f does not change a sign:

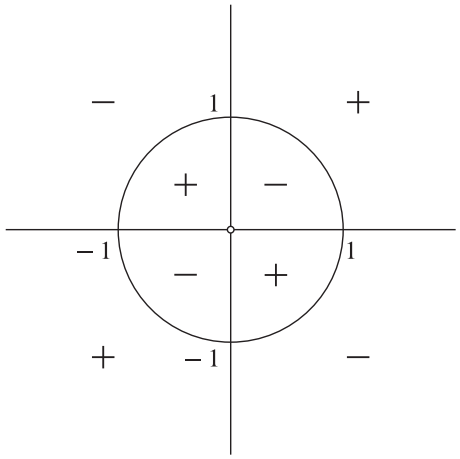


FIGURE 6.

Now it is obvious, that the function does not have a local extrem at any of the point $[1,0]$, $[0,1]$, $[-1,0]$ and $[0,-1]$ – each of them has a function value equal to zero, but in any neigbourhood of any of them there are points with both positive and negative function values.

We shall now prove, that f has at points $[-1/\sqrt{2e}, 1/\sqrt{2e}]$ a $[1/\sqrt{2e}, -1/\sqrt{2e}]$ local maximum and at the points $[1/\sqrt{2e}, 1/\sqrt{2e}]$ a $[-1/\sqrt{2e}, -1/\sqrt{2e}]$ má f local minimum. Define the function $\bar{f}\colon \mathbb{R}^2 \to \mathbb{R}$ this way:

$$\bar{f}(x,y) = \begin{cases} f(x,y) & \text{pro } [x,y] \in D_f, \\ 0 & \text{pro } [x,y] = [0,0]. \end{cases}$$

Take for example the point $\vec{a} = [1/\sqrt{2e}, 1/\sqrt{2e}]$. This point lies in the interior of the set $J = \{[x,y] \in \mathbb{R}^2;\ x \geq 0,\ y \geq 0,\ x^2 + y^2 \leq 1\}$. The set J is compact, the function $\bar{f}$ is continuous on it and thus has there a maximum and a minimum.

At each point of the set $H(J)$ the function $\bar{f}$ attains the value zero, which is its maximum on J (since it is non-positive on J). Since the function $\bar{f}$ is negative at the points of $\operatorname{Int} J$, it has a minimum at some interior point of the set J. From the previous part of the solution we know, that the only suspicious point is $\vec{a}$. The function $\bar{f}$ has thus a minimum at the point $\vec{a}$ on the set J, and since $\vec{a} \in \operatorname{Int} J$, then $\bar{f}$ has a local minimum at the point $\vec{a}$. Since $f = \bar{f}$ on the neighbourhood of the point $\vec{a}$, then also the function f has a local minimum at the point $\vec{a}$. ♣

Example 38. Find local extrems of the function $f(x,y) = x + 2y + \frac{3}{4}x^2 + xy + 2y^2$ on the set $M = \{[x,y] \in \mathbb{R}^2;\ y^2 - 2 \leq x \leq -y^2 + 2\}$.

Solution. Draw a figure of the set M.

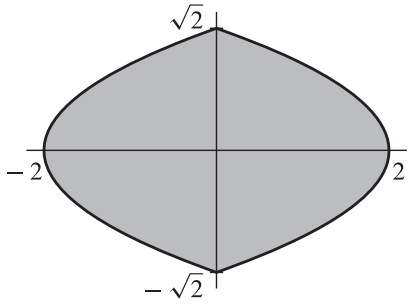


FIGURE 7.

The set M is compact.Since f is continuous on M, it has a maximum and a minimum on M. Examine firstly suspicious points in the interior of the setp M, i.e. solve the linear system

$$1 + \frac{3}{2}x + y = 0, \qquad 2 + x + 4y = 0.$$

Only one point makes this linear system valid: $[x, y] = [-2/5, -2/5] \in \operatorname{Int} M$.

We can rewrite the boundary $H(M)$ in the form $H(M) = M_1 \cup M_2$, kde

$$M_1 = \{[x, y] \in \mathbb{R}^2;\ x = y^2 - 2, |y| \leq \sqrt{2}\},$$
$$M_2 = \{[x, y] \in \mathbb{R}^2;\ x = 2 - y^2, |y| \leq \sqrt{2}\}.$$

Find points suspicious from being an extrem on the set M_1. We define supporting funcion φ by:

$$\varphi(y) = f(y^2 - 2, y) = \frac{3}{4}y^4 + y^3 + 1, \quad y \in \mathbb{R}.$$

We are interested in extremes of the function φ on the set $[-\sqrt{2}, \sqrt{2}]$. We have $\varphi'(y) = 3y^2(y+1)$, and thus $\varphi'(y) = 0$ for just two points: $y = 0$ a $y = -1$. These points together with the points $-\sqrt{2}$ and $\sqrt{2}$ gives us the following points suspicious from being an extrem on M_1 for the function f: $[0, \sqrt{2}]$, $[0, -\sqrt{2}]$, $[-2, 0]$, $[-1, -1]$.

We take a similar approach on the set M_2. In this case we examine a supporting function

$$\psi(y) = f(2 - y^2, y) = \frac{3}{4}y^4 - y^3 - 2y^2 + 4y + 5$$

and we solve the equation $\psi'(y) = 3y^3 - 3y^2 - 4y + 4 = 0$ on $(-\sqrt{2}, \sqrt{2})$. One root of the equation is 1, and thus we can write $\psi'(y) = (y - 1)(3y^2 - 4)$. The

second bracket is equal to zero for $y = 2/\sqrt{3}$ and $y = -2/\sqrt{3}$. Hence we get another suspicious points $[1,1]$, $[2/3, 2/\sqrt{3}]$ and $[2/3, -2/\sqrt{3}]$.

since we know, that f has a maximum and minimum on the set M it is sufficient to compare the function values at the suspicious points:

$$\begin{aligned} f(-2/5,-2/5) &= -3/5, & f(-2,0) &= 1,\\ f(-1,-1) &= 3/4, & f(0,\sqrt{2}) &= 4+2\sqrt{2},\\ f(0,-\sqrt{2}) &= 4-2\sqrt{2}, & f(1,1) &= 27/4,\\ f(2/3,2/\sqrt{3}) &= \frac{16+11\sqrt{3}}{3\sqrt{3}}, & f(2/3,-2/\sqrt{3}) &= \frac{11\sqrt{3}-16}{3\sqrt{3}}. \end{aligned}$$

Hence we easily find out, that the function f attains a maximum value at the point $[0,\sqrt{2}]$ with the function value $\max\limits_{M} f = 4+2\sqrt{2}$, a minimum at the point $[-2/5,-2/5]$ with the function value $\min\limits_{M} f = -3/5$. ♣

1.4. Implicit function theorem

Consider an equation

$$x^2 + y^2 - 1 = 0. \tag{12}$$

Our task is to solve the equation (12) for y in terms of the parameter x. The equation has a solution only for $x \in [-1,1]$, and then $y = \sqrt{1-x^2}$ and $y = -\sqrt{1-x^2}$. Dependent on x x we have two (for $x \in (-1,1)$), or one solution (for $x = \pm 1$), or no solution (for $x \notin [-1,1]$). We can see, that for $x \in (-1,1)$ is it not possible to calculate from the equation (12) exactly one y dependent on x. But if we use restriction to an appropriate (that is enough small) neighbourhood U of the point $x_0 \in (-1,1)$ and a neighbourhood V of the point y_0, where x_0, y_0 satisfy the equation (12), then it is possible to find exactly one $y \in V$ for each $x \in U$ such that x and y satisfy the equation (12). The equation (12) thus define on the neighbourhood U some function φ of one real variable x with the values in the neighbourhood V, which satisfies $x^2 + (\varphi(x))^2 - 1 = 0$, for each $x \in U$. The situation is maybe shown better on the following figure.

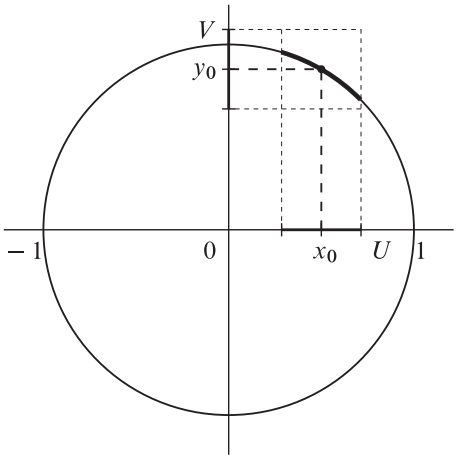


FIGURE 8.

Notice, that for points $x_0 = 1$ a $x_0 = -1$ we can not find neighbourhoods U and V with stated property.

The following theorem is generally dealing with the situation illustrated in the previous example. It describes, what condition do we need to get exactly one variable y as the function of the variable x from the equation $F(x, y) = 0$.

Theorem 39 (Implicit function theorem)**.** Let $G \subset \mathbb{R}^{n+1}$ be open, $F\colon G \to \mathbb{R}$, $\tilde{\vec{x}} = [\tilde{x}_1, \ldots, \tilde{x}_n] \in \mathbb{R}^n$, $\tilde{y} \in \mathbb{R}$, $[\tilde{\vec{x}}, \tilde{y}] \in G$ and let:

(i) $F \in \mathcal{C}^1(G)$,

(ii) $F(\tilde{\vec{x}}, \tilde{y}) = 0$,

(iii) $\dfrac{\partial F}{\partial y}(\tilde{\vec{x}}, \tilde{y}) \neq 0$.

Then there exist neighbourhood $U \subset \mathbb{R}^n$ of the point $\tilde{\vec{x}}$ and neighbourhood $V \subset \mathbb{R}$ of the point $\tilde{y}$ such that for each $\vec{x} \in U$ there exists exactly one $y \in V$ satisfying $F(\vec{x}, y) = 0$. If we denote this y by the symbol $\varphi(\vec{x})$, then $\varphi \in \mathcal{C}^1(U)$ and

$$\frac{\partial \varphi}{\partial x_j}(\vec{x}) = -\frac{\frac{\partial F}{\partial x_j}\big(\vec{x}, \varphi(\vec{x})\big)}{\frac{\partial F}{\partial y}\big(\vec{x}, \varphi(\vec{x})\big)}, \quad \text{provided } j \in \{1, \ldots, n\}, \vec{x} \in U. \tag{13}$$

Proof. We prove only the first part of the theorem for $n = 1$. We can assume without loss of generality, that $\frac{\partial F}{\partial y}(\tilde{x}, \tilde{y}) > 0$. Since G is an opne set, $F \in \mathcal{C}^1(G)$ and $\frac{\partial F}{\partial y}(\tilde{x}, \tilde{y}) > 0$, then there exist $\delta_1 > 0$ and $\eta > 0$ such that

$$\forall [x, y] \in [\tilde{x} - \delta_1, \tilde{x} + \delta_1] \times [\tilde{y} - \eta, \tilde{y} + \eta]\colon \frac{\partial F}{\partial y}(x, y) > 0. \tag{14}$$

The function $t \mapsto F(\tilde{x}, t)$ is due to (14) is increasing in the interval $[\tilde{y} - \eta, \tilde{y} + \eta]$. From that and from the $F(\tilde{x}, \tilde{y}) = 0$ we obtain $F(\tilde{x}, \tilde{y} + \eta) > 0$ and $F(\tilde{x}, \tilde{y} - \eta) < 0$. The continuity of the function F imply the existence of $\delta_2 \in (0, \delta_1)$ such that

$$\forall x \in (\tilde{x} - \delta_2, \tilde{x} + \delta_2) \colon (F(x, \tilde{y} + \eta) > 0 \ \ \& \ \ F(x, \tilde{y} - \eta) < 0).$$

Set $U = (\tilde{x} - \delta_2, \tilde{x} + \delta_2)$ and $V = (\tilde{y} - \eta, \tilde{y} + \eta)$. Choose $x \in U$. The function of one variable $t \mapsto F(x, t)$ is increasing and continuous on the interval $[\tilde{y} - \eta, \tilde{y} + \eta]$ and $F(x, \tilde{y} + \eta) > 0$, $F(x, \tilde{y} - \eta) < 0$ holds. The function thus attains on this interval all values between $F(x, \tilde{y} + \eta)$ and $F(x, \tilde{y} - \eta)$ (theorem **??**), and each of them at exacly one point. Taht means that there exists exactly one $y \in V$ satisfying $F(x, y) = 0$.

The proof of the assertion that $\varphi \in \mathcal{C}^1(U)$ is somewhat more difficult and we do not give it here. But we will show, how to derive the formula (13) provided that, we already know that $\varphi \in \mathcal{C}^1(U)$. For each $x \in U$ $F(x, \varphi(x)) = 0$ holds. It is equality of two functions on the neighbourhood U (the function $x \mapsto F(x, \varphi(x))$ and the function $x \mapsto 0$), from which follows also equality of their derivatives on the neigbourhood U. According to the Theorem 32 $\frac{\partial F}{\partial x}\big(x, \varphi(x)\big) \cdot 1 + \frac{\partial F}{\partial y}\big(x, \varphi(x)\big)\varphi'(x) = 0$ holds. From that the formula (13) follows, because $\frac{\partial F}{\partial y}\big(x, \varphi(x)\big) \neq 0$ for $x \in U$. ∎

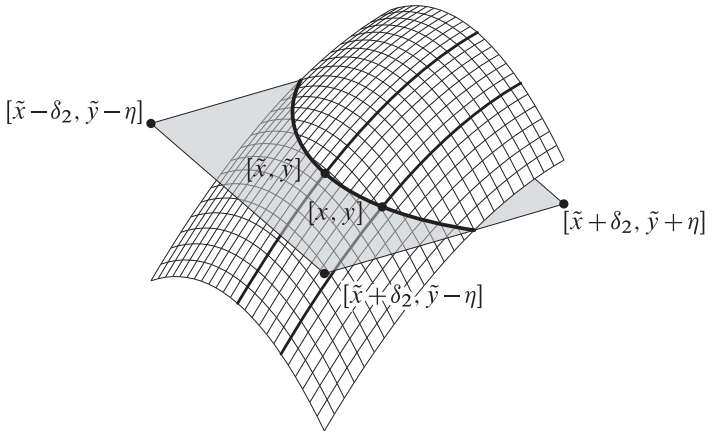


FIGURE 9.

Remark. It can be proved, that the function φ is as "smooth" as the function F is. Let F be for example of the class $\mathcal{C}^\infty$, then also φ is of the class$\mathcal{C}^\infty$.

Example 40. Let $M = \{[x, y] \in \mathbb{R}^2;\ (x^2 + y^2)^2 - 2(x^2 - y^2) = 0\}$. Show that in a neighbourhood of the point $[\sqrt{3}/2, 1/2]$ is it possible to describe the set M as a graph of s function φ of the variable x. Calculate $\varphi'(\sqrt{3}/2)$.

Solution. Set

$$F(x, y) = (x^2 + y^2)^2 - 2(x^2 - y^2).$$

Then:

(i) $F \in \mathcal{C}^1(\mathbb{R}^2)$,
(ii) $F(\sqrt{3}/2, 1/2) = 0$,
(iii) $\frac{\partial F}{\partial y}(\sqrt{3}/2, 1/2) = \big(2(x^2 + y^2) \cdot 2y + 4y\big)\Big|_{[\sqrt{3}/2,1/2]} = 4 \neq 0.$[5]

The assumption of the implicit function theorem are then satisfied. According to this theorem follows that the set M is described by the function φ in a neigbourhood of the point $[\sqrt{3}/2, 1/2]$. Calculate the derivative of the function φ at the point $\sqrt{3}/2$. We get

$$\frac{\partial F}{\partial x}(\sqrt{3}/2, 1/2) = \big(2(x^2 + y^2) \cdot 2x - 4x\big)\Big|_{[\sqrt{3}/2,1/2]} = 0,$$

and thus due to (13)

$$\varphi'(\sqrt{3}/2) = -\frac{\frac{\partial F}{\partial x}(\sqrt{3}/2, 1/2)}{\frac{\partial F}{\partial y}(\sqrt{3}/2, 1/2)} = -\frac{0}{4} = 0.$$

On the first two figures we can see parts of the graph of the function F and on the third figure, there is a set M.

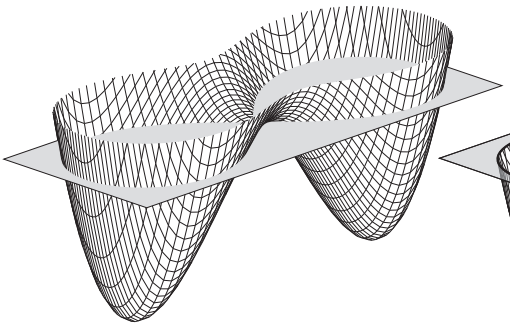


FIGURE 10.

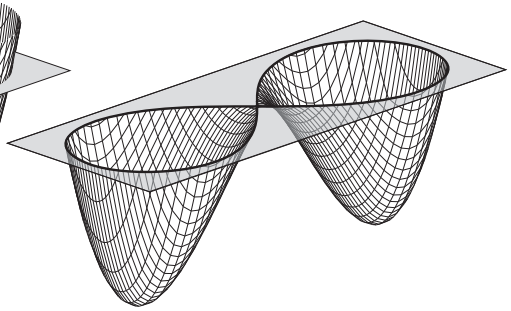


FIGURE 11.

[5]The symbol $\big(2(x^2 + y^2) \cdot 2y + 4y\big)\big|_{[\sqrt{3}/2,1/2]}$ denotes the value of the expression $\big(2(x^2 + y^2) \cdot 2y + 4y\big)$ at the point $[\sqrt{3}/2, 1/2]$. We will use this notation also in the following lecture notes.

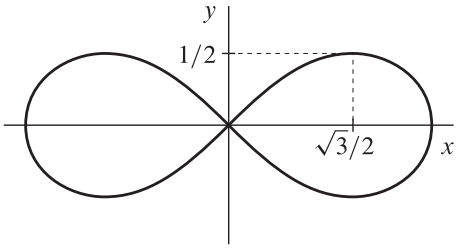


FIGURE 12.

♣

We introduce a more general theorem without the proof.

Theorem 41 (Implicit functions theorem)**.** Let $n, m \in \mathbb{N}$, $G \subset \mathbb{R}^{n+m}$ be an open set, $F_j \colon G \to \mathbb{R}$, $j = 1, \ldots, m$, $\tilde{\vec{x}} \in \mathbb{R}^n$, $\tilde{\vec{y}} \in \mathbb{R}^m$ satisfying$[\tilde{\vec{x}}, \tilde{\vec{y}}] = [\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{y}_1, \ldots, \tilde{y}_m] \in G$ and then:

(i) $F_j \in \mathcal{C}^1(G)$ for $j \in \{1, \ldots, m\}$,

(ii) $F_j(\tilde{\vec{x}}, \tilde{\vec{y}}) = 0$ pro $j \in \{1, \ldots, m\}$, that is

$$\begin{aligned} F_1(\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{y}_1, \ldots, \tilde{y}_m) &= 0, \\ &\vdots \\ F_m(\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{y}_1, \ldots, \tilde{y}_m) &= 0, \end{aligned}$$

(iii) and finally

$$\begin{vmatrix} \frac{\partial F_1}{\partial y_1}(\tilde{\vec{x}}, \tilde{\vec{y}}) & \ldots & \frac{\partial F_1}{\partial y_m}(\tilde{\vec{x}}, \tilde{\vec{y}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1}(\tilde{\vec{x}}, \tilde{\vec{y}}) & \ldots & \frac{\partial F_m}{\partial y_m}(\tilde{\vec{x}}, \tilde{\vec{y}}) \end{vmatrix} \neq 0.$$

Then we can find a neigbourhood $U \subset \mathbb{R}^n$ of the point $\tilde{\vec{x}}$ and a neigbourhood $V \subset \mathbb{R}^m$ of the point $\tilde{\vec{y}}$ such that for each $\vec{x} \in U$ there exist exactly one $\vec{y} \in V$ with the property$F_j(\vec{x}, \vec{y}) = 0$ for each $j \in \{1, \ldots, m\}$. If we denote coordinates of this $\vec{y}$ as $\varphi_j(\vec{x})$, $j = 1, \ldots, m$, then $\varphi_j \in \mathcal{C}^1(U)$.

Remarks. 1. In the condition (iii) of the previous theorem, there appeared a symbol, which is non-defined yet. It is so-called *determinant*. Its exact definition and basic properties are stated in a section 2.3. For $m = 2$ and $a, b, c, d \in \mathbb{R}$ this holds:

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

For $m = 1$ is the condition (iii) in a form $\frac{\partial F_1}{\partial y_1}(\tilde{\vec{x}}, \tilde{y}) \neq 0$, and then the Theorem 39 is a special case of the Theorem 41.

2. If the functions $F_1, \ldots, F_m$ from the Theorem 41 are of the class $\mathcal{C}^\infty(G)$, it can be even proved, that also functions $\varphi_1, \ldots, \varphi_m$ will be of the class $\mathcal{C}^\infty(U)$. We will use this assertion in the following examples and exercises.

3. If we consider the linear system

$$\begin{aligned} F_1(x_1, \ldots, x_n, y_1, \ldots, y_m) &= 0, \\ &\vdots \\ F_m(x_1, \ldots, x_n, y_1, \ldots, y_m) &= 0 \end{aligned}$$

of m equation with n real parameters $x_1, \ldots, x_n$ and m unknowns $y_1, \ldots, y_m$, then the Implicit functions theorem gives us - besides other things - conditions, when we can "calculate" the unknowns $y_1, \ldots, y_m$ from this linear system dependent on the parameters $x_1, \ldots, x_n$.

Example 42. Show that the set

$$\big\{[x, y] \in \mathbb{R}^2;\ e^{xy} - \sin(x + y) = 1\big\}$$

is (on some neighbourhood of the point $[0, -\pi]$) a graph of the function $x \mapsto y(x)$ of the class $\mathcal{C}^\infty$, which satisfies $y(0) = -\pi$. Calculate $y'(0)$ a $y''(0)$.

Solution. Denote $F(x, y) = e^{xy} - \sin(x + y) - 1$ and check, that the function satisfy assumptions of the Implicit function theorem at the point $[0, -\pi]$:

(i) $F \in \mathcal{C}^\infty(\mathbb{R}^2)$,
(ii) $F(0, -\pi) = e^0 - \sin(-\pi) - 1 = 1 - 0 - 1 = 0$,
(iii) $\dfrac{\partial F}{\partial y}(0, -\pi) = xe^{xy} - \cos(x + y)\Big|_{[0,-\pi]} = 1 \neq 0.$

Thus there exist numbers $\delta_1 > 0$, $\delta_2 > 0$ such that

$$\forall x \in (-\delta_1, \delta_1)\ \exists! y(x) \in (-\pi - \delta_2, -\pi + \delta_2)\colon F\big(x, y(x)\big) = 0$$

and the function $x \mapsto y(x)$ is of the class $\mathcal{C}^\infty$ on $(-\delta_1, \delta_1)$. The function

$$x \mapsto e^{xy(x)} - \sin\big(x + y(x)\big) - 1$$

is thus on the interval $(-\delta_1, \delta_1)$ constant, and hence its derivation

$$x \mapsto e^{xy(x)}\big(y(x) + xy'(x)\big) - \cos\big(x + y(x)\big)\big(1 + y'(x)\big) \tag{15}$$

is on the interval $(-\delta_1, \delta_1)$ equal to 0. Setting $x = 0$ $y(0) = -\pi$, we obtain easily that $y'(0) = \pi - 1$.

We use the function(15) to calculate the second derivative. As we determined, is the function constant on the interval $(-\delta_1, \delta_1)$ as well, and thus its derivative is equal to zero on this interval. Hence,

$$e^{xy(x)}\big(y(x)+xy'(x)\big)^2 + e^{xy(x)}\big(2y'(x)+xy''(x)\big)+ \\ +\sin\big(x+y(x)\big)\big(1+y'(x)\big)^2 - \cos\big(x+y(x)\big)y''(x)=0.$$

Setiing $x=0$ and $y(0)=-\pi$ and $y'(0)=\pi-1$ we obtain $y''(0)=-\pi^2-2\pi+2$. ♣

Example 43. Show that the set

$$\{[x,y,z]\in\mathbb{R}^3;\ x\sin z + y\cos z - \exp z = 0\}$$

is (on some neighbourhood of the point $[2,1,0]$) a graph of the function $[x,y]\mapsto z(x,y)$, which satisfies $z(2,1)=0$. Write the equation of a tangent plane (if it exists) to the graph of the function z at the point $[2,1,0]$.

Solution. Denote $F(x,y,z)=x\sin z+y\cos z-\exp z$ and check the assumption of the Implicit function theorem for F at the point $[2,1,0]$.

(i) $F\in\mathcal{C}^1(\mathbb{R}^3)$,
(ii) $F(2,1,0)=2\cdot 0+1\cdot 1-1=0$,
(iii) $\dfrac{\partial F}{\partial z}(2,1,0)=(x\cos z-y\sin z-\exp z)\big|_{[2,1,0]}=1\neq 0.$

The assumption of the Theorem 39 are satisfied, thus there exist numbers $\delta_1>0$, $\delta_2>0$ such that

$$\forall[x,y]\in B([2,1],\delta_1)\ \exists!z(x,y)\in(-\delta_2,\delta_2)\colon F\big(x,y,z(x,y)\big)=0$$

and the function $[x,y]\mapsto z(x,y)$ is of the class $\mathcal{C}^1$ on $B([2,1],\delta_1)$. Thus there exist a tangent plane to the graph of the function z at the point $[2,1,0]$. The function

$$[x,y]\mapsto F\big(x,y,z(x,y)\big)=x\sin z(x,y)+y\cos z(x,y)-\exp z(x,y)$$

is constant on the ball $B([2,1],\delta_1)$, and thus its partial derivative are equal to zero there. We obtain

$$\sin z(x,y)+x\cos z(x,y)\cdot\frac{\partial z}{\partial x}(x,y)- \\ -y\sin z(x,y)\cdot\frac{\partial z}{\partial x}(x,y)-\exp z(x,y)\cdot\frac{\partial z}{\partial x}(x,y)=0,$$

$$x\cos z(x,y)\cdot\frac{\partial z}{\partial y}(x,y)+\cos z(x,y)- \\ -y\sin z(x,y)\cdot\frac{\partial z}{\partial y}(x,y)-\exp z(x,y)\cdot\frac{\partial z}{\partial y}(x,y)=0.$$

After setting $[x,y]=[2,1]$ and $z(2,1)=0$ we get

$$\frac{\partial z}{\partial x}(2,1)=0, \qquad \frac{\partial z}{\partial y}(2,1)=-1.$$

The second way, how to calculate these partial derivative is using the formula (13). The tangent plane to the graph of the function $[x,y]\mapsto z(x,y)$ at the point $[2,1,0]$ is described by the function

$$T(x,y)=0\cdot(x-2)-1\cdot(y-1)=1-y.$$

♣

Example 44. Prove that there exist functions $[x,y]\mapsto u(x,y)$, $[x,y]\mapsto v(x,y)$ of the class $\mathcal{C}^\infty$ satisfying $u(1,1)=0$, $v(1,1)=\pi/4$ and relations

$$\exp\left(\frac{u}{x}\right)\cos\left(\frac{v}{y}\right)=\frac{x}{\sqrt{2}}, \qquad \exp\left(\frac{u}{x}\right)\sin\left(\frac{v}{y}\right)=\frac{y}{\sqrt{2}}$$

on some neighbourhood of the point $[1,1]$. Find a tangent plane to the graph of the function u (v, respectively) at the point $[1,1,0]$ ($[1,1,\pi/4]$, respectively).

Solution. Denote

$$F_1(x,y,u,v)=\exp\left(\frac{u}{x}\right)\cos\left(\frac{v}{y}\right)-\frac{x}{\sqrt{2}},$$
$$F_2(x,y,u,v)=\exp\left(\frac{u}{x}\right)\sin\left(\frac{v}{y}\right)-\frac{y}{\sqrt{2}}.$$

Check that the function F_1 and F_2 satisfy the assumption of the Implicit functions theorem (Theorem 41) at the point $[1,1,0,\pi/4]$.

(i) $F_1,F_2\in\mathcal{C}^1(G)$, where $G=(0,+\infty)^2\times\mathbb{R}^2$,

(ii) $F_1(1,1,0,\pi/4)=\cos(\pi/4)-1/\sqrt{2}=0$,
$F_2(1,1,0,\pi/4)=\sin(\pi/4)-1/\sqrt{2}=0$,

(iii)

$$\begin{vmatrix} \frac{\partial F_1}{\partial u}(1,1,0,\pi/4) & \frac{\partial F_1}{\partial v}(1,1,0,\pi/4) \\ \frac{\partial F_2}{\partial u}(1,1,0,\pi/4) & \frac{\partial F_2}{\partial v}(1,1,0,\pi/4) \end{vmatrix} =$$

$$= \begin{vmatrix} \frac{1}{x}\exp(\frac{u}{x})\cos(\frac{v}{y}) & -\frac{1}{y}\exp(\frac{u}{x})\sin(\frac{v}{y}) \\ \frac{1}{x}\exp(\frac{u}{x})\sin(\frac{v}{y}) & \frac{1}{y}\exp(\frac{u}{x})\cos(\frac{v}{y}) \end{vmatrix}_{[1,1,0,\pi/4]} = 1\neq 0 \text{ (see the section 2.3).}$$

Thus there exist posive numbers $\delta_1>0$, $\delta_2>0$ such that

$$\forall[x,y]\in B([1,1],\delta_1)\ \exists!\,[u(x,y),v(x,y)]\in B([0,\pi/4],\delta_2):$$
$$F_1\big(x,y,u(x,y),v(x,y)\big)=0\ \ \&\ \ F_2\big(x,y,u(x,y),v(x,y)\big)=0,$$

the functions u, v are of the class $\mathcal{C}^1$ on $B([1,1],\delta_1)$ and $u(1,1)=0, v(1,1)=\pi/4$ holds. By calculating the derivatives of the contant functions

$$[x,y]\mapsto F_1\big(x,y,u(x,y),v(x,y)\big),$$
$$[x,y]\mapsto F_2\big(x,y,u(x,y),v(x,y)\big)$$

on $B([1,1],\delta_1)$ with respect to x we obtain

$$\exp\left(\frac{u(x,y)}{x}\right)\frac{\frac{\partial u}{\partial x}(x,y)\cdot x-u(x,y)}{x^2}\cos\left(\frac{v(x,y)}{y}\right)-$$
$$-\exp\left(\frac{u(x,y)}{x}\right)\sin\left(\frac{v(x,y)}{y}\right)\frac{1}{y}\frac{\partial v}{\partial x}(x,y)-\frac{1}{\sqrt{2}}=0,$$

$$\exp\left(\frac{u(x,y)}{x}\right)\frac{\frac{\partial u}{\partial x}(x,y)\cdot x-u(x,y)}{x^2}\sin\left(\frac{v(x,y)}{y}\right)+$$
$$+\exp\left(\frac{u(x,y)}{x}\right)\cos\left(\frac{v(x,y)}{y}\right)\frac{1}{y}\frac{\partial v}{\partial x}(x,y)=0.$$

Similarly by calculating the derivative with respect to y we obtain

$$\exp\left(\frac{u}{x}\right)\frac{1}{x}\frac{\partial u}{\partial y}\cos\left(\frac{v}{y}\right)-\exp\left(\frac{u}{x}\right)\sin\left(\frac{v}{y}\right)\frac{\frac{\partial v}{\partial y}y-v}{y^2}=0,$$
$$\exp\left(\frac{u}{x}\right)\frac{1}{x}\frac{\partial u}{\partial y}\sin\left(\frac{v}{y}\right)+\exp\left(\frac{u}{x}\right)\cos\left(\frac{v}{y}\right)\frac{\frac{\partial v}{\partial y}y-v}{y^2}-\frac{1}{\sqrt{2}}=0.$$

Here we used an useful notation, which makes the writing clearer – we omitted arguments x a y of the functions u, v. Setting $[x,y]=[1,1]$, we obtain two systems of linear equation

$$\frac{\partial u}{\partial x}(1,1)-\frac{\partial v}{\partial x}(1,1)=1,\qquad \frac{\partial u}{\partial x}(1,1)+\frac{\partial v}{\partial x}(1,1)=0$$

and

$$\frac{\partial u}{\partial y}(1,1)-\frac{\partial v}{\partial y}(1,1)=-\frac{\pi}{4},\qquad \frac{\partial u}{\partial y}(1,1)+\frac{\partial v}{\partial y}(1,1)=1+\frac{\pi}{4}.$$

Hence,

$$\frac{\partial u}{\partial x}(1,1)=\frac{1}{2},\qquad \frac{\partial v}{\partial x}(1,1)=-\frac{1}{2},\qquad \frac{\partial u}{\partial y}(1,1)=\frac{1}{2},\qquad \frac{\partial v}{\partial y}(1,1)=\frac{\pi}{4}+\frac{1}{2}.$$

From the continuity of the partial derivatives of the function u (v, respectively) at the point $[1,1]$ it follows the existence of a tangent plane at the point $[1,1,0]$

($[1,1,\pi/4]$, respectively). The tangent plane is described by the function

$$T(x,y)=\frac{1}{2}(x-1)+\frac{1}{2}(y-1),\ \text{resp. } T(x,y)=\frac{\pi}{4}-\frac{1}{2}(x-1)+\left(\frac{1}{2}+\frac{\pi}{4}\right)(y-1).$$

♣

1.5. Lagrange multipliers theorem

The following theorem describes a method for a function of the class $\mathcal{C}^1$, how to find points suspicious for being an extrem on a set, which is a contour line of a function of the class $\mathcal{C}^1$.

Theorem 45 (Lagrange multiplier theorem)**.** Let $G\subset\mathbb{R}^2$ be an open set, $f,g\in\mathcal{C}^1(G)$, $M=\{[x,y]\in G;\ g(x,y)=0\}$ and $[\tilde{x},\tilde{y}]\in M$ be a local extrem point of the function f on the set M. Then at least one of the following conditions is satisfied:

(i) $\nabla g(\tilde{x},\tilde{y})=\vec{o}$,

(ii) there exists a real number $\lambda\in\mathbb{R}$ satisfyign

$$\frac{\partial f}{\partial x}(\tilde{x},\tilde{y})+\lambda\frac{\partial g}{\partial x}(\tilde{x},\tilde{y})=0, \tag{16}$$

$$\frac{\partial f}{\partial y}(\tilde{x},\tilde{y})+\lambda\frac{\partial g}{\partial y}(\tilde{x},\tilde{y})=0. \tag{17}$$

Proof. It is sufficient to prove that if (i) is not satisfied, then the second condition (ii) holds. Suppose then, that $\nabla g(\tilde{x},\tilde{y})\neq\vec{o}$. We can assume without loss of generality that $\frac{\partial g}{\partial y}(\tilde{x},\tilde{y})\neq 0$. If this partial derivative was equal to zero, then it would have to be $\frac{\partial g}{\partial x}(\tilde{x},\tilde{y})\neq 0$ and the whole following procedure will be the same except for changing the roles of x and y.

Let $\delta\in\mathbb{R}$, $\delta>0$, be such that f attains at the point $[\tilde{x},\tilde{y}]$ extrem value on $B([\tilde{x},\tilde{y}],\delta)\cap M$. Set $G'=B([\tilde{x},\tilde{y}],\delta)$. According to the Implicit function theorem used on the set G', funkction $g|_{G'}$ and a point $[\tilde{x},\tilde{y}]$ there exist neighbourhood U of the point $\tilde{x}$, neighbourhood V of the point $\tilde{y}$ aand function $\varphi\colon U\to V$ of the class $\mathcal{C}^1$ satisfying

- $G'\cap M\cap(U\times V)=\operatorname{graf}\varphi$,
- $\varphi(\tilde{x})=\tilde{y}$.

Define a function $h\colon U\to\mathbb{R}$ by $h(x)=f\big(x,\varphi(x)\big)$. According to the Theorem 32, the function h is of the class $\mathcal{C}^1$ on U, since $f\in\mathcal{C}^1(G)$ and $\varphi\in\mathcal{C}^1(U)$. If f has a maximum at the point $[\tilde{x},\tilde{y}]$ on $G'\cap M$, then for each $x\in U$ follows

$$h(x)=f\big(x,\varphi(x)\big)\leq f(\tilde{x},\tilde{y})=f\big(\tilde{x},\varphi(\tilde{x})\big)=h(\tilde{x}),$$

since $[x, \varphi(x)] \in G' \cap M$. The function h has thus a maximum at the point $\tilde{x}$ on U. If f has a minimum at the point $[\tilde{x}, \tilde{y}]$ on $G' \cap M$, then we could derive similarly that h has a minimum at the point$\tilde{x}$ on U.

since the function h has a local extrem at the point $\tilde{x}$, $h'(\tilde{x}) = 0$ must hold. According to the Theorem 32 then follows

$$h'(\tilde{x}) = \frac{\partial f}{\partial x}\big(\tilde{x}, \varphi(\tilde{x})\big) + \frac{\partial f}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)\varphi'(\tilde{x}) = 0. \tag{18}$$

From the Implicit function theorem we know that

$$\varphi'(\tilde{x}) = -\frac{\frac{\partial g}{\partial x}\big(\tilde{x}, \varphi(\tilde{x})\big)}{\frac{\partial g}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)}. \tag{19}$$

Seting

$$\lambda = -\frac{\frac{\partial f}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)}{\frac{\partial g}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)},$$

then the relation (17) is automatically satisfied. If we use (19) and (18), we get

$$\begin{aligned}\frac{\partial f}{\partial x}(\tilde{x}, \tilde{y}) + \lambda\frac{\partial g}{\partial x}(\tilde{x}, \tilde{y}) &= \frac{\partial f}{\partial x}(\tilde{x}, \tilde{y}) - \frac{\frac{\partial f}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)}{\frac{\partial g}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)} \cdot \frac{\partial g}{\partial x}(\tilde{x}, \tilde{y}) = \\ &= \frac{\partial f}{\partial x}(\tilde{x}, \tilde{y}) + \frac{\partial f}{\partial y}\big(\tilde{x}, \varphi(\tilde{x})\big)\varphi'(\tilde{x}) = 0.\end{aligned}$$

That completed the proof. ∎

Example 46. Find a maximum and a minimum of the function $f(x, y) = 2x + 8y$ on the set $M = \{[x, y] \in \mathbb{R}^2;\ x^2 + 2y^2 = 1\}$.

Solution. The set M is compact and the function f is continuous, it thus has a maximum and a minimum on M. To find points, where could be a local extrem of the function f on the set M, we use the previous theorem. Set $G = \mathbb{R}^2$ and $g(x, y) = x^2 + 2y^2 - 1$. The function f and g are of the class $\mathcal{C}^1$ and $\frac{\partial f}{\partial x}(x, y) = 2$, $\frac{\partial f}{\partial x}(x, y) = 8$, $\frac{\partial g}{\partial x}(x, y) = 2x$ a $\frac{\partial f}{\partial x}(x, y) = 4y$ holds. then $\nabla g(x, y) = \vec{o}$, if and only if $[x, y] = [0, 0]$, but the set M does not contain this point. at any point of the set M the condition (i) from the Theorem 45 is not satisfied, at the points of the local extremes of the function f on the set M there must be satisfied the second condition (ii). Thus, to find a points suspicious from being an extrem, we have to solve the following linear system:

$$\begin{aligned} 2 + 2\lambda x &= 0, \\ 8 + 4\lambda y &= 0, \\ x^2 + 2y^2 &= 1. \end{aligned} \tag{20}$$

The first two equation are the condition (ii) from the Theorem 45, the last equation express, that we are finding points which are contained in the set M. If we multiply the first equation by the number 4 and subtract from it the second equation od ní druhou rovnici, we get $4\lambda(2x - y) = 0$ must hold. From the first equation we can see that $\lambda \neq 0$, thus $y = 2x$. After setting to the last equation we obtain that for the solution of the liear system $x = \frac{1}{3}$, $y = \frac{2}{3}$, nebo $x = -\frac{1}{3}$, $y = -\frac{2}{3}$. In the first case, we get $\lambda = -3$, in the second then $\lambda = 3$. Because $f(\frac{1}{3}, \frac{2}{3}) = 1$ and $f(-\frac{1}{3}, -\frac{2}{3}) = -1$, the function f attains a maximum value on M at the point $[\frac{1}{3}, \frac{2}{3}]$ and a minimum on M at the point $[-\frac{1}{3}, -\frac{2}{3}]$.

Notice, that finally there was no need to calculate the value λ. It was enough, that for values x, y other than $[\frac{1}{3}, \frac{2}{3}]$ or $[-\frac{1}{3}, -\frac{2}{3}]$ the linear system (20) has no solution. Usually, it is easier to set all found pairs $[x, y]$ to the function f in this moment, to reduce their number by searching the value of λ. ♣

Without the proof, we introduce a more general form of the Theorem 45, where the set M is described by several conditions. This formulation uses a notion linear dependance of the vectors. This notion will be defined in the section 2.2. Here we will only notice, that one vector is linear dependent if and only if it is a zero vector, and two vectors are linear dependent if and only if one of them is a multiple of the other.

Theorem 47 (Lagrange multipliers theorem)**.** Let $m, n \in \mathbb{N}$, $m < n$, $G \subset \mathbb{R}^n$ be an open set, $f, g_1, \ldots, g_m \in \mathcal{C}^1(G)$,

$$M = \{\vec{z} \in G;\ g_1(\vec{z}) = 0, g_2(\vec{z}) = 0, \ldots, g_m(\vec{z}) = 0\}$$

and the point $\vec{\tilde{z}} \in M$ is a point of a local extrem of the function f on the set M. Then at least one of the following condition is satisfied:

(i) vectors $\nabla g_1(\vec{\tilde{z}}), \nabla g_2(\vec{\tilde{z}}), \ldots, \nabla g_m(\vec{\tilde{z}})$ are linear dependent,

(ii) There exist real numbers $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}$ satisfying

$$\nabla f(\vec{\tilde{z}}) + \lambda_1 \nabla g_1(\vec{\tilde{z}}) + \lambda_2 \nabla g_2(\vec{\tilde{z}}) + \cdots + \lambda_m \nabla g_m(\vec{\tilde{z}}) = \vec{o}.$$

Remarks. 1. Setting $m = 1$ a $n = 2$, we obtain from the Theorem 47 the Theorem 45.

2. We call the numbers $\lambda_1, \ldots, \lambda_m$ **multipliers**.

Example 48. Find a maximum and a minimum of the function $f(x, y, z) = xyz$ on the set $M = \{[x, y, z] \in \mathbb{R}^3;\ x^2 + y^2 + z^2 = 1, x + y + z = 0\}$.

Solution. The set M is compact and the function f is continuous, it thus has a maximum and a minimum on M. We use the previous theorem to find points, where could be a local extrem of the function f on the set M. Set $G = \mathbb{R}^3$,

$$g_1(x, y, z) = x^2 + y^2 + z^2 - 1, \qquad g_2(x, y, z) = x + y + z.$$

Functions f, g_1 and g_2 are of the class $\mathcal{C}^1$. Calculate partial derivatives.

$$\frac{\partial f}{\partial x}(x,y,z) = yz, \qquad \frac{\partial f}{\partial y}(x,y,z) = xz, \qquad \frac{\partial f}{\partial z}(x,y,z) = xy,$$
$$\frac{\partial g_1}{\partial x}(x,y,z) = 2x, \qquad \frac{\partial g_1}{\partial y}(x,y,z) = 2y, \qquad \frac{\partial g_1}{\partial z}(x,y,z) = 2z,$$
$$\frac{\partial g_2}{\partial x}(x,y,z) = 1, \qquad \frac{\partial g_2}{\partial y}(x,y,z) = 1, \qquad \frac{\partial g_2}{\partial z}(x,y,z) = 1.$$

The vectors $[2x, 2y, 2z]$ and $[1, 1, 1]$ are linear dependent if and only of $x = y = z$ holds. There is no point with this property contained in the set M, since for the point $[x, x, x]$ there must be $g_1(x, x, x) = 3x^2 - 1 = 0$ and $g_2(x, x, x) = 3x = 0$ at the same time, which is not possible. Thus it is necessary to solve this nonlinear system:

$$yz + \lambda_1 2x + \lambda_2 = 0, \tag{21}$$
$$xz + \lambda_1 2y + \lambda_2 = 0, \tag{22}$$
$$xy + \lambda_1 2z + \lambda_2 = 0, \tag{23}$$
$$x^2 + y^2 + z^2 - 1 = 0, \tag{24}$$
$$x + y + z = 0. \tag{25}$$

By subtracting (22) from (21) we obtain:

$$-z(x-y) + 2\lambda_1(x-y) = 0. \tag{26}$$

Hence it follows, that must be $z = 2\lambda_1$ or $x = y$. Similarly by subtracting (23) from (22) we get:

$$-x(y-z) + 2\lambda_1(y-z) = 0. \tag{27}$$

That gives us $x = 2\lambda_1$ or $y = z$. From the relations (26) and (27) thus follows, that must be either $x = y$, or $y = z$, or $x = z$. Look at the first case, where $x = y$. From (25) we have $z = -2x$ and from (24) we get $6x^2 = 1$, i.e. $x = 1/\sqrt{6}$ or $x = -1/\sqrt{6}$. Indeed, we can calculate corresponding y, z, λ_1 and λ_2 to this points. We can solve similarly the other cases $y = z$ and $z = x$. We obtain these suspicious points:

$$\left[\tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right], \quad \left[-\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right], \quad \left[-\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}\right],$$
$$\left[-\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right], \quad \left[\tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right], \quad \left[\tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}\right].$$

Setting the values to the function f in these points we find out that in the first row, there are maxima points of the function f on M, and in the second row, there are minima points of f on M. ♣

Example 49. Find a maximum and a minimum of the function $f(x, y, z) = xyz$ on the set $M = \{[x, y, z] \in \mathbb{R}^3;\ x^2 + y^2 + z^2 \leq 1, x + y + z \geq 0\}$.

Solution. The set M is compact, since M is a closed half ball. The function f is continuous on M, therefore it has a maximum and a minimum on M. We will search the points suspicious of being an extrem separately on the interior of the set M and on the boundary of M. Apart of these point there cannot be another extrem point, since if $\vec{x} \in A \subset M$ a extrem point on M, then it is a extrem point on A.

The interior of M is equal to $\{[x,y,z] \in \mathbb{R}^3;\ x^2+y^2+z^2 < 1, x+y+z > 0\}$. The function f is of the class $\mathcal{C}^1$. The suspicious points on $\operatorname{Int} M$ are points with all first partial derivative equal to 0. It is $\nabla f(x,y,z) = [yz, xz, xy]$. This vector is equal to zero vector at point with at leat two zero coordinates, that is on coordinate axis. The suspicious points on $\operatorname{Int} M$ are points from any of the following sets:

$$\{[x,0,0];\ x \in (0,1)\}, \quad \{[0,y,0];\ y \in (0,1)\} \quad \text{a} \quad \{[0,0,z];\ z \in (0,1)\}.$$

We divide the boundary $H(M)$ into parts

$$\begin{aligned} H_1 &= \{[x,y,z] \in G_1;\ x+y+z = 0\}, \quad \text{where} \\ G_1 &= \{[x,y,z] \in \mathbb{R}^3;\ x^2+y^2+z^2 < 1\}, \\ H_2 &= \{[x,y,z] \in G_2;\ x^2+y^2+z^2 = 1\}, \quad \text{where} \\ G_2 &= \{[x,y,z] \in \mathbb{R}^3;\ x+y+z > 0\}, \\ H_3 &= \{[x,y,z] \in \mathbb{R}^3;\ x^2+y^2+z^2 = 1, x+y+z = 0\}. \end{aligned}$$

Notice that the sets G_1 and G_2 are open. We can use the Lagrange multipliers theorem to find suspicious points on the set H_1, H_2, H_3, respectively.

The function $[x,y,z] \mapsto x+y+z$ has non-zero gradient on $\mathbb{R}^3$, therefore in the case of the set H_1 we get the suspicious points by solving the system of equations

$$\begin{aligned} yz + \lambda &= 0, \\ xz + \lambda &= 0, \\ xy + \lambda &= 0, \\ x+y+z &= 0. \end{aligned}$$

By a similar procedure as in the previous example we get the only solution of this system $[x,y,z] = [0,0,0]$. This point is contained in H_1, and thus it is a point, which is suspicious of being an extrem.

In the case of the set H_2 has the function $[x,y,z] \mapsto x^2+y^2+z^2-1$ zero gradient only at the point $[0,0,0]$, which is not an element of H_2, therefore we get

the suspicious points by solving the system of equation

$$\begin{aligned} yz + 2\lambda x &= 0, \\ xz + 2\lambda y &= 0, \\ xy + 2\lambda z &= 0, \\ x^2 + y^2 + z^2 &= 1. \end{aligned}$$

By a similar procedure to the one in the previous example we get the solution, and there we omit the corresponding values of λ:

$$\left[-\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}\right], \left[-\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \left[\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \left[\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}\right],$$
$$\left[-\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \left[-\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}\right], \left[\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}\right], \left[\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right],$$
$$[1,0,0], \quad [-1,0,0], \quad [0,1,0], \quad [0,-1,0], \quad [0,0,1], \quad [0,0,-1].$$

From all these points, only these points are contained in H_2:

$$\left[\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \quad \left[-\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \quad \left[\tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}\right], \quad \left[\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{3}}, -\tfrac{1}{\sqrt{3}}\right],$$
$$[1,0,0], \quad [0,1,0], \quad [0,0,1].$$

We examined the set H_3 already in the previous example. Here are the suspicious points

$$\left[\tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right], \quad \left[-\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}\right], \quad \left[-\tfrac{1}{\sqrt{6}}, -\tfrac{1}{\sqrt{6}}, \tfrac{2}{\sqrt{6}}\right],$$
$$\left[-\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right], \quad \left[\tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}\right], \quad \left[\tfrac{1}{\sqrt{6}}, \tfrac{1}{\sqrt{6}}, -\tfrac{2}{\sqrt{6}}\right].$$

Comparing the values of the function f in the suspicious points we get that the function f has a maximum on M at the point $\left[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$ and a minimum on M at the point $\left[-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right], \left[\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right], \left[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right]$. ♣

Example 50. Find extremes of the function $f(x,y,z) = xy + z^2$ on the set $M = \{[x,y,z] \in \mathbb{R}^3;\ x^2 + y^2 + z^2 \leq 1,\ x \geq 0\}$.

Solution. The set M is compact, since M is closed half ball. The function f is continuous on M, therefore it has a maximum and minimum on M. We will proceed similarly to previous example. We will search the points suspicious of being an extrem separately on the interior of the set M and on the boundary of M. It can be easily saw, that the only point with all partial derivative equal to zero is $[0,0,0]$. But this point is not an element of $\operatorname{Int} M$.

Examine the function f on a part of the boundary

$$H_1 = \{[x,y,z] \in \mathbb{R}^3;\ x^2 + y^2 + z^2 = 1, x > 0\}.$$

We want to use the Lagrange theorem. Set $G = \{[x, y, z] \in \mathbb{R}^3;\ x > 0\}$,

$$g(x, y, z) = x^2 + y^2 + z^2 - 1$$

and calculate

$$\frac{\partial g}{\partial x} = 2x, \quad \frac{\partial g}{\partial y} = 2y, \quad \frac{\partial g}{\partial z} = 2z.$$

It is obvious, that at each point of the set H_1 is at at least one of these partial derivative non-zero. The functions f and g are of the class$\mathcal{C}^1$ on $\mathbb{R}^3$, thus we can use the Lagrange multipliers theorem. The points, suspicious of being an extrem, solve this system of equations

$$\begin{aligned} y + 2\lambda x &= 0, \\ x + 2\lambda y &= 0, \\ 2z + 2\lambda z &= 0, \\ x^2 + y^2 + z^2 - 1 &= 0. \end{aligned}$$

The third equation is satisfied, if eiter $z = 0$, or $\lambda = -1$ holds.

a) The case $z = 0$. Multiply the first equation by y, the second by $-x$ and add them. We get the equation $y^2 - x^2 = 0$. Setting this to the fourth equation, we obtain $2x^2 = 1$. It is either $x = 1/\sqrt{2}$, or $x = -1/\sqrt{2}$. But the value $x = -1/\sqrt{2}$ does not satisfy the condition $x > 0$. We get the first pair of suspicious points: $[1/\sqrt{2}, 1/\sqrt{2}, 0]$ a $[1/\sqrt{2}, -1/\sqrt{2}, 0]$.

b) If $\lambda = -1$, then the first two equation has the only one solution $x = y = 0$. This solution does not satisfy the condition $x > 0$.

The second part of the boundary is the set $H_2 = \{[0, y, z] \in \mathbb{R}^3;\ y^2 + z^2 \leq 1\}$ (a disc in the plane $x = 0$). Often it is handy to use a special shape of the function f or of the set, where we examine the function instead of a general procedure using Lagrange multipliers theorem. It is so also in this case. We define a supporting function

$$\varphi(y, z) = f(0, y, z) = z^2$$

and find extremes of the function φ of two variables on the compact set $L = \{[y, z] \in \mathbb{R}^2;\ y^2 + z^2 \leq 1\}$, and thus also extremes of the function f on the set H_2. If the function f does not have an extrem on the set H_2 at a point of H_2, then it also does not have an extrem on M at this point. Into the set of points suspicious of being an extrem of f on M we thus add only extrem points of f on H_2.

The function φ is non-negative and has zero values on L just at points of the set $K = \{[y, 0] \in \mathbb{R}^2;\ y \in [-1, 1]\}$. The function φ thus has a minimum on L at some point of K.

On L, there holds $\varphi(y, z) = z^2 \leq y^2 + z^2 \leq 1$ and φ attains the value 1 on L just at points $[0, 1]$ and $[0, -1]$ and therefore they are the maxima points of φ on L.

Now it is enough to compare the function values in all suspicious points. We obtain:

$$f(1/\sqrt{2}, 1/\sqrt{2}, 0) = 1/2, \qquad f(0,0,-1) = f(0,0,1) = 1,$$
$$f(1/\sqrt{2}, -1/\sqrt{2}, 0) = -1/2, \qquad f(0,y,0) = 0 \text{ for } y \in [-1,1].$$

Thus it is $\max\limits_{M} f = 1$ and $\min\limits_{M} f = -1/2$. ♣

Example 51. Examine the extremes of the function $f(x,y) = (x+y)\exp(-x^2 - y^2)$ on the set $M = \{[x,y] \in \mathbb{R}^2;\ x^2 + y^2 \le 1, |x| \le y + 1\}$.

Solution. Draw the set M.

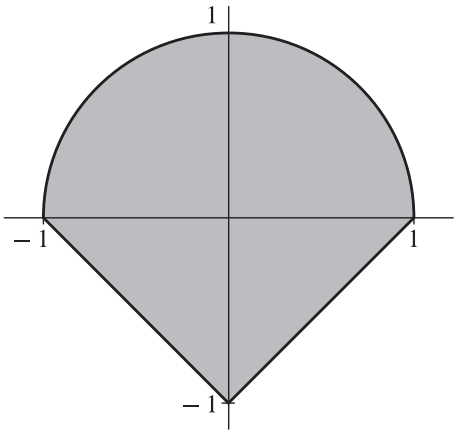


FIGURE 13.

The set M is compact and the function f is continuous on it. The function f thus has a maximum and minimum on M. Find suspicious points in $\mathrm{Int}\, M$. Solve a system of equations on $\mathrm{Int}\, M$

$$\frac{\partial f}{\partial x} = \exp(-x^2 - y^2)(1 - 2x^2 - 2xy) = 0,$$
$$\frac{\partial f}{\partial y} = \exp(-x^2 - y^2)(1 - 2y^2 - 2xy) = 0.$$

This system has the only one solution in $\mathrm{Int}\, M$ and it is $[1/2, 1/2]$.

Next find suspicious points on the boundary $H(M)$. Firstly consider this part of the boundary

$$H_1 = \{[x,y] \in \mathbb{R}^2;\ 0 \le x \le 1, y = x - 1\}.$$

Define a supporting function

$$\varphi(x) = f(x, x-1) = (2x-1)\exp(-2x^2+2x-1).$$

The function φ is continuous on a compact set $[0,1]$, and thus has there a maximum and a minimum. Since φ' has non-zero values at all points of the interval $(0,1)$, then the only suspicious points are the endpoints $x=0$, $x=1$. There are two suspicious points on H_1, namely $[0,-1]$ and $[1,0]$.

Proceed similarly on this part of the boundary

$$H_2 = \{[x,y] \in \mathbb{R}^2;\ -1 \le x \le 0, y = -x-1\}$$

we get the following suspicious points $[-1/2,-1/2]$ and $[-1,0]$.

We use the multiplier theorem on this part of the boundary

$$H_3 = \{[x,y] \in \mathbb{R}^2;\ x^2+y^2 = 1, y > 0\}$$

Denote

$$g(x,y) = x^2+y^2-1$$

and calculate $\frac{\partial g}{\partial x}(x,y) = 2x$, $\frac{\partial g}{\partial y}(x,y) = 2y$. For all points from H_3 is second of the partial derivative non-zero. The functions f and g are of the class $\mathcal{C}^1$ on the whole $\mathbb{R}^2$. The assumption of the multiplier theorem are thus satisfied. Suspicious will be the points, which solve the system of equations on H_3

$$\begin{aligned}
\exp(-x^2-y^2)(1-2x^2-2xy)+2\lambda x &= 0,\\
\exp(-x^2-y^2)(1-2y^2-2xy)+2\lambda y &= 0,\\
x^2+y^2-1 &= 0.
\end{aligned}$$

We multiply the first equation by y, the second by $-x$ and then we add them together. Then we obtain $\exp(-x^2-y^2)(y-x)=0$, which is satisfied if and only if $x=y$. From the third equation we can see that this situation could happen H_3 at the only point $[x,y] = [1/\sqrt{2}, 1/\sqrt{2}]$.

Since we know, that the function f has extremes on M, it is enough to compare the function values in suspicious points:

$$\begin{aligned}
f(1/2,1/2) &= \exp(-1/2), & f(0,-1) &= -\exp(-1),\\
f(1,0) &= \exp(-1), & f(-1/2,-1/2) &= -\exp(-1/2),\\
f(-1,0) &= -\exp(-1), & f(1/\sqrt{2},1/\sqrt{2}) &= \sqrt{2}\exp(-1).
\end{aligned}$$

We get then

$$\begin{aligned}
\max_M f &= f(1/2,1/2) = \exp(-1/2),\\
\min_M f &= f(-1/2,-1/2) = -\exp(-1/2).
\end{aligned}$$

♣

1.6. Concave and quasiconcave functions

In this section we will study concave and quasiconcave functions of several variables. The definition of a concave multivariate function is a straightforward generalization of the notion concave function of one variable from the chapter **??**.

Definition. Let $M \subset \mathbb{R}^n$. We say that M is a **convex set** if:

$$\forall \vec{x}, \vec{y} \in M \; \forall t \in [0,1] \colon t\vec{x} + (1-t)\vec{y} \in M.$$

Remarks. 1. The set M is convex if and only if every line segment with endpoints in M lies whole in M.

2. If the sets $M, N \subset \mathbb{R}^n$ are convex, then also a set $M \cap N$ is convex. But a set $M \cup N$ does not have to be convex in general.

Definition. Let $M \subset \mathbb{R}^n$ be a convex set and a function f be defined on M. We say that f is a **concave function** on M if:

$$\forall \vec{a}, \vec{b} \in M \; \forall t \in [0,1] \colon f(t\vec{a} + (1-t)\vec{b}) \geq tf(\vec{a}) + (1-t)f(\vec{b}).$$

We say that f is **strictly concave function** on M if:

$$\forall \vec{a}, \vec{b} \in M, \vec{a} \neq \vec{b} \; \forall t \in (0,1) \colon f(t\vec{a} + (1-t)\vec{b}) > tf(\vec{a}) + (1-t)f(\vec{b}).$$

Remark. If we reverse an inequality sign, we get a definition of **convex** and **strictly convex** function of several variables.

Remark. The concave function does not have to be continuous on its domain, it can be shown by a function f defined on interval $[0,1]$ by

$$f(x) = \begin{cases} 0 & \text{pro } x = 0, \\ 1 & \text{pro } x \in (0,1]. \end{cases}$$

However, the following theorem holds; we omit its somewhat more difficult proof.

Theorem 52. Let $G \subset \mathbb{R}^n$ be an open convex set and a function f be concave on G. Then f is continuous on G.

Theorem 53. Let a function f be concave on a convex set M. Then for each $\alpha \in \mathbb{R}$ is a set $Q_\alpha = \{\vec{x} \in M;\ f(\vec{x}) \geq \alpha\}$ convex.

Proof. Let $\alpha \in \mathbb{R}$. If $\vec{a}, \vec{b} \in Q_\alpha$ and $t \in [0,1]$, then $f(\vec{a}) \geq \alpha$ and $f(\vec{b}) \geq \alpha$. From that and from the concavity of the function f follows

$$f(t\vec{a} + (1-t)\vec{b}) \geq tf(\vec{a}) + (1-t)f(\vec{b}) \geq t\alpha + (1-t)\alpha = \alpha,$$

in other words $t\vec{a} + (1-t)\vec{b} \in Q_\alpha$. ∎

The following theorem says, that for functions of the class $\mathcal{C}^1$ is the concavity of the function f on the set G is equal to the property that the graph of the function f lies under every tangent hyperplane to the graph of the function.

Theorem 54. Let $G \subset \mathbb{R}^n$ be a convex open set and $f \in \mathcal{C}^1(G)$. Then the function f is concave on G if and only if

$$\forall \vec{x}, \vec{y} \in G\colon\ f(\vec{y}) \leq f(\vec{x}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{x})(y_i - x_i). \tag{28}$$

Proof. $\Rightarrow$ For each $\vec{x}, \vec{y} \in G$ and for each $t \in (0, 1]$

$$f\big((1-t)\vec{x} + t\vec{y}\big) \geq (1-t)f(\vec{x}) + tf(\vec{y})$$

holds, and thus

$$\frac{f\big(\vec{x} + t(\vec{y} - \vec{x})\big) - f(\vec{x})}{t} \geq f(\vec{y}) - f(\vec{x}).$$

From the theorem about composite function derivative and from the theorem about limit and order (Theorem **??**) it follows

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{x})(y_i - x_i) = \lim_{t \to 0+} \frac{f\big(\vec{x} + t(\vec{y} - \vec{x})\big) - f(\vec{x})}{t} \geq f(\vec{y}) - f(\vec{x}).$$

$\Leftarrow$ Choose $\vec{v}, \vec{w} \in G$ and next $t \in [0, 1]$. Apply (28) on pairs of points $t\vec{v} + (1-t)\vec{w}$, $\vec{v}$:

$$f(\vec{v}) - f(t\vec{v} + (1-t)\vec{w}) \leq \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(t\vec{v} + (1-t)\vec{w})(v_i - w_i)(1-t);$$

and also on pairs of points $t\vec{v} + (1-t)\vec{w}$, $\vec{w}$:

$$f(\vec{w}) - f(t\vec{v} + (1-t)\vec{w}) \leq \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(t\vec{v} + (1-t)\vec{w})(w_i - v_i)t.$$

Multiply the first inequality by t and second by $(1-t)$ and then add them together:

$$tf(\vec{v}) - tf(t\vec{v} + (1-t)\vec{w}) + (1-t)f(\vec{w}) - (1-t)f(t\vec{v} + (1-t)\vec{w}) \leq 0.$$

We alter the previous inequality and get

$$tf(\vec{v}) + (1-t)f(\vec{w}) \leq f(t\vec{v} + (1-t)\vec{w}).$$

This would complete the proof. ∎

From the previous theorem we could easily derive the following assertion.

Theorem 55. Let $G \subset \mathbb{R}^n$ be a convex open set and $f \in \mathcal{C}^1(G)$ be concave on G. If $\vec{a} \in G$ stationary point of the function f, then $\vec{a}$ is a maximum point of the function f on the set G.

Proof. From the Theorem 54 we get, that for each point $\vec{y} \in G$ follows

$$f(\vec{y}) \leq f(\vec{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a})(y_i - a_i) = f(\vec{a}) + \sum_{i=1}^{n} 0 \cdot (y_i - a_i) = f(\vec{a}),$$

which completes the proof. ∎

We state - without the proof - the following theorem, which characterize a strict concavity of the function.

Theorem 56. Let $G \subset \mathbb{R}^n$ be a convex open set and $f \in \mathcal{C}^1(G)$. Then the function f is strictly concave on G if and only if the following expression holds

$$\forall \vec{x}, \vec{y} \in G, \vec{x} \neq \vec{y}\colon f(\vec{y}) < f(\vec{x}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{x})(y_i - x_i).$$

Definition. Let $M \subset \mathbb{R}^n$ be a convex set and f be a function defined on M. We say that f is **quasiconcave** on M, if

$$\forall \vec{a}, \vec{b} \in M\ \forall t \in [0,1]\colon f(t\vec{a} + (1-t)\vec{b}) \geq \min\{f(\vec{a}), f(\vec{b})\}.$$

We say that f is **strictly quasiconcave** on M, if

$$\forall \vec{a}, \vec{b} \in M, \vec{a} \neq \vec{b}\ \forall t \in (0,1)\colon f(t\vec{a} + (1-t)\vec{b}) > \min\{f(\vec{a}), f(\vec{b})\}.$$

In this figure, there is a quasiconcave function, which is not concave.

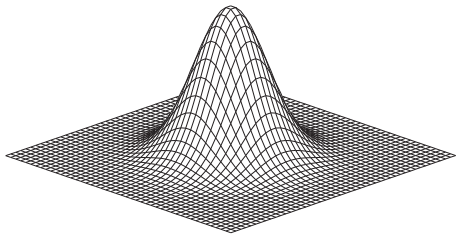


FIGURE 14.

Remark. If we write in the previous definition $f(t\vec{a}+(1-t)\vec{b}) \leq \max\{f(\vec{a}), f(\vec{b})\}$ instead of $f(t\vec{a} + (1-t)\vec{b}) \geq \min\{f(\vec{a}), f(\vec{b})\}$, we obtain a definition of a **quasiconvex** function of several variables. Similarly we can define a **strictly quasiconvex** function of several variables.

In this section we deal only with (strictly) concave and (strictly) quasiconcave funkctions. But we can alter the formulation of the stated results straightforwardly also for convex and quasiconvex functions. It is enough to realize, that the function f is convex if and only if the function $-f$ is concave, and similarly it holds for function strictly convex, quasiconvex and strictly quasiconvex.

Remark. It can be shown, that strictly quasiconcave functions are exactly the quasiconcave functions, whose graph "does not contain a horizontal line segment", i.e.

$$\operatorname{non}\big(\exists \vec{a}, \vec{b} \in M, \vec{a} \neq \vec{b},\ \forall t \in [0,1]\colon\ f(t\vec{a} + (1-t)\vec{b}) = f(\vec{a})\big).$$

Theorem 57. Let $M \subset \mathbb{R}^n$ be a convex set and f be a function defined on M. Then:

(i) If f is concave on M, then it is also quasiconcave M.
(ii) If f is strictly concave on M, then it is also strictly quasiconcave on M.

Proof. (i) Take $\vec{a}, \vec{b} \in M$ and $t \in [0,1]$. We can assume without loss of generality, that $f(\vec{b}) \geq f(\vec{a})$. Then

$$f(t\vec{a} + (1-t)\vec{b}) \geq tf(\vec{a}) + (1-t)f(\vec{b}) \geq tf(\vec{a}) + (1-t)f(\vec{a}) = f(\vec{a}).$$

(ii) Take $\vec{a}, \vec{b} \in M$, $\vec{a} \neq \vec{b}$, and $t \in (0,1)$. We can assume without loss of generality, that $f(\vec{b}) \geq f(\vec{a})$. Then

$$f(t\vec{a} + (1-t)\vec{b}) > tf(\vec{a}) + (1-t)f(\vec{b}) \geq tf(\vec{a}) + (1-t)f(\vec{a}) = f(\vec{a}).$$

■

Theorem 58. Let $M \subset \mathbb{R}^n$ be a convex set and f be a function defined on M. The function f is quasiconcave on M if and only if for each $\alpha \in \mathbb{R}$ the set $Q_\alpha = \{\vec{x} \in M;\ f(\vec{x}) \geq \alpha\}$ is convex.

Proof. $\Rightarrow$ Let $\vec{a}, \vec{b} \in Q_\alpha$, $\alpha \in \mathbb{R}$, $t \in [0,1]$. Then

$$f(t\vec{a} + (1-t)\vec{b}) \geq \min\{f(\vec{a}), f(\vec{b})\} \geq \alpha,$$

and hence $t\vec{a} + (1-t)\vec{b} \in Q_\alpha$.

$\Leftarrow$ Let $\vec{a}, \vec{b} \in M$ and $t \in [0,1]$. Denote $\alpha = \min\{f(\vec{a}), f(\vec{b})\}$. Then $\vec{a}, \vec{b} \in Q_\alpha$ and a set Q_α is convex, hence $t\vec{a} + (1-t)\vec{b} \in Q_\alpha$. From it follows that

$$f(t\vec{a} + (1-t)\vec{b}) \geq \alpha = \min\{f(\vec{a}), f(\vec{b})\}.$$

■

Compare the previous theorem with the Theorem 53.

The last two theorems of this section are important, because they express the relation of a strict quasiconcavity of the function and uniqueness of a maximum point.

Theorem 59. Let $M \subset \mathbb{R}^n$ be a convex set and f is strictly quasiconcave function on M. If f has a maximum on M , then it has exactly one maximum point.

Proof. Let $\vec{a}, \vec{b} \in M$ be two different points, at which f attains its maximum value on M. If we use condition of strict quasiconcavity for $\vec{a}, \vec{b}$ and $t = 1/2$, we get $f\left(\frac{1}{2}\vec{a} + \frac{1}{2}\vec{b}\right) > f(\vec{a}) = \max_M f$ ant that is a contradiction. ∎

Theorem 60. Let $M \subset \mathbb{R}^n$ be a convex, bounded, closed and non-empty set and f be continuous and strictly quasiconcave function on M. Then f has a maximum on M at exactly one point.

Proof. The function f has a maximum on M since M is non-empty compact set and f is continuous on M. According to the previous theorem, the uniqueness then follows from the strict quasiconcavity. ∎

1.7. Exercise

In the following five exercises examine, if the given set M is open, closed or bounded and determine its boundary, interior and closure.

1. $\{[x, y] \in \mathbb{R}^2;\ 1 \leq x < 2,\ 1 \leq y < 2\}$

2. $\left\{[x, y] \in \mathbb{R}^2;\ \left|\frac{y-1}{x}\right| \leq 1\right\}$

3. $\left\{[1/n, 1/m] \in \mathbb{R}^2;\ n \in \mathbb{N},\ m \in \mathbb{N}\right\}$

4. $\left\{[x, y] \in \mathbb{R}^2;\ \frac{4-4x^2-y^2}{4y} \geq 0\right\}$

5. $\left\{[3\cos t + \cos 3t, 3\sin t - \sin 3t] \in \mathbb{R}^2;\ t \in [0, 2\pi]\right\}$

6. Let for each $k \in \mathbb{N}$ je $M_k = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \leq (1 - 1/k)^2\}$. Define $M = \bigcup_{k=1}^{\infty} M_k$.[6] Determine, if the sets M, M_k are open or closed.

7. Prove the continuity of the function defined by

$$f(x, y) = \begin{cases} (x+y)^2 \sin\left(\frac{1}{\sqrt{x^2+y^2}}\right) & \text{for } [x, y] \neq [0, 0], \\ 0 & \text{for } [x, y] = [0, 0], \end{cases}$$

on the whole domain.

In the following exercises examine for a given function f a domain, continuity, contour lines and determine a maximum and a minimum, if they exist.

8. $f(x, y) = x^2 - y^2$ **9.** $f(x, y) = x/y$

10. $f(x, y) = \arcsin xy$

[6] The symbol $\bigcup_{k=1}^{\infty}$ means the same as $\bigcup_{k \in \mathbb{N}}$.

11. $f(x,y) = \begin{cases} \frac{2x^2y}{x^4+y^2} & \text{for } [x,y] \neq [0,0] \\ 0 & \text{for } [x,y] = [0,0] \end{cases}$

12. Determine a domain of the function $f(x,y) = \sqrt{x^2+4y^2}+1$. Calculate partial derivatives at all points, where they exist.

13. For the function

$$f(x,y) = \begin{cases} \frac{xy}{\sqrt{x^2+y^2}} & \text{pro } [x,y] \neq [0,0], \\ 0 & \text{pro } [x,y] = [0,0] \end{cases}$$

calculate partial derivatives at all points, where they exist.

14. Determine a domain of the function $f(x,y,z) = (x/y)^z$; calculate first and second partial derivatives at all points, where they exist. Determine an equation of a tangent plane to the graph of the function f at the point $[e,1,2,e^2]$.

15. Determine a domain of a function $f(x,y) = \operatorname{arctg} \frac{x-y}{x+y}$. calculate first and second partial derivatives and write an equation of a tangent plane to the graph of the function f at the point $[1,1,f(1,1)]$.

16. Examine extremes of the function $f(x,y) = xy\exp(-xy)$ on the set

$$M = \{[x,y] \in \mathbb{R}^2;\ x \geq 0, y \geq 0\}.$$

17. Examine extremes of the function $f(x,y) = \dfrac{x-y}{1+x^2+y^2}$ on the set

$$M = \{[x,y] \in \mathbb{R}^2;\ y \geq 0\}.$$

18. Examine extremes of the function $f(x,y) = xy(1-x-y)$ on the set

$$M = \{[x,y] \in \mathbb{R}^2;\ x \geq 0, y \geq 0, x+y \leq 1\}.$$

19. Examine extremes of the function $f(x,y) = \sin x \cos y$ on the set

$$M = \{[x,y] \in \mathbb{R}^2;\ 0 \leq x \leq 2\pi, 0 \leq y \leq -x+2\pi\}.$$

20. Examine extremes of the function $f(x,y,z) = (x-1)^2+(2y-1)^2+(z-2)^2$ on the set

$$M = \{[x,y,z] \in \mathbb{R}^3;\ x \geq 0, y \geq 0, z \geq 0, x+y+z \leq 4\}.$$

21. Prove that the set

$$\{[x,y] \in \mathbb{R}^2;\ y = x + \log y\}$$

is a graph of the function $x \mapsto y(x)$ at a neighbourhood of the point $[e-1,e]$. Write the equation of a tangent (if it exists) to the graph of the function y at the point $[e-1,e]$.

22. Prove that the set

$$M = \{[x, y] \in \mathbb{R}^2;\ x^2 + 2xy + y^2 - 4x + 2y - 2 = 0\}$$

is a graph of the function $x \mapsto y(x)$ at a neighbourhood of the point $[1, 1]$. Examine, if the function y is convex or concave on a neighbourhood of the point 1.

23. Prove that there exist functions $x \mapsto y(x)$ and $x \mapsto z(x)$ of the class $\mathcal{C}^\infty$, which satisfies $y(1) = e$, $z(1) = 1$, and relations

$$\exp z - xyz = 0, \quad \log(xy) - \frac{x}{z} = 0$$

on a neighbourhood of the point 1. Calculate $y'(1)$ and $z'(1)$.

In the following exercise prove that the set M is a graph of the function $[x, y] \mapsto z(x, y)$ of the class $\mathcal{C}^\infty$ at a neighbourhood of a given point. Determine a tangent plane to the graph of the function at a given point.

24. $M = \{[x, y, z] \in \mathbb{R}^3;\ x^2 + 2y^2 + 3z^2 + xy - z - 9 = 0\}, \quad [1, -2, 1]$

25. $M = \{[x, y, z] \in \mathbb{R}^3;\ \exp z + x^2 y + z + 5 = 0\}, \quad [1, -6, 0]$

26. $M = \{[x, y, z] \in \mathbb{R}^3;\ \cos^2 x + \cos^2 y + \cos^2 z = 1\}, \quad [\pi/3, \pi/2, \pi/6]$

In the following exercises find extremes of the function f non the set M.

27. $f(x, y) = 25x^3 - 18xy + 9x$, $M = \{[x, y] \in \mathbb{R}^2;\ x \in [-3, 3],\ y^2 - 5x^2 \le 4\}$

28. $f(x, y) = 6xy + y^3 + 6y$, $M = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \le 5\}$

29. $f(x, y) = xy + 2x + 3y$, $M = \{[x, y] \in \mathbb{R}^2;\ 4x^2 + 9y^2 \le 36\}$

30. $f(x, y) = xy + 2x + 3y$, $M = \{[x, y] \in \mathbb{R}^2;\ 4x^2 + 9y^2 \le 36,\ y \le -x/2\}$

31. $f(x, y) = \exp(x^2 - y^2 + y)$, $M = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 \le 1,\ y \ge 0\}$

32. $f(x, y, z) = x^2 + y^2 + z^2$, $M = \{[x, y, z] \in \mathbb{R}^3;\ x^2 + y^2 = z^2,\ x - 2z = 3\}$

Results of exercises

1. M is neither open, nor closed; it is bounded; $H(M) = \{[x, 1] \in \mathbb{R}^2;\ 1 \le x \le 2\} \cup \{[x, 2] \in \mathbb{R}^2;\ 1 \le x \le 2\} \cup \{[1, y] \in \mathbb{R}^2;\ 1 \le y \le 2\} \cup \{[2, y] \in \mathbb{R}^2;\ 1 \le y \le 2\}$; $\operatorname{Int} M = \{[x, y] \in \mathbb{R}^2;\ 1 < x < 2,\ 1 < y < 2\}$; $\overline{M} = \{[x, y] \in \mathbb{R}^2;\ 1 \le x \le 2,\ 1 \le y \le 2\}$

2. M is neither open, nor closed; it is not bounded; $H(M) = \{[x, y] \in \mathbb{R}^2;\ y = -x + 1\} \cup \{[x, y] \in \mathbb{R}^2;\ y = x + 1\}$; $\operatorname{Int} M = \{[x, y] \in \mathbb{R}^2;\ \left|\frac{y-1}{x}\right| < 1\}$; $\overline{M} = M \cup \{[0, 1]\}$

3. M is neither open, nor closed; it is bounded; $H(M) = M \cup \{[1/n, 0] \in \mathbb{R}^2;\ n \in \mathbb{N}\} \cup \{[0, 1/m] \in \mathbb{R}^2;\ m \in \mathbb{N}\} \cup \{[0, 0]\}$; $\operatorname{Int} M = \emptyset$, $\overline{M} = H(M)$

4. M is neither open, nor closed; it is not bounded; $H(M) = \{[x, y] \in \mathbb{R}^2;\ 4 - 4x^2 - y^2 = 0\} \cup \{[x, 0] \in \mathbb{R}^2;\ x \in \mathbb{R}\}$; $\operatorname{Int} M = \{[x, y] \in \mathbb{R}^2;\ \frac{4-4x^2-y^2}{4y} > 0\}$; $\overline{M} = M \cup \{[x, 0] \in \mathbb{R}^2;\ x \in \mathbb{R}\}$

5. M is closed, bounded; $H(M) = \overline{M} = M$; $\operatorname{Int} M = \emptyset$

6. For each $k \in \mathbb{N}$ the set M_k is closed and not open; the set $M = \{[x, y] \in \mathbb{R}^2;\ x^2 + y^2 < 1\}$ is open and not closed.

8. $D_f = \mathbb{R}^2$; f is continuous on D_f; $f_{-1}(\{0\}) = \{[x, y] \in \mathbb{R}^2;\ y = x\} \cup \{[x, y] \in \mathbb{R}^2;\ y = -x\}$, $f_{-1}(\{1\}) = \{[x, y] \in \mathbb{R}^2;\ x^2 - y^2 = 1\}$, $f_{-1}(\{-1\}) = \{[x, y] \in \mathbb{R}^2;\ x^2 - y^2 = -1\}$; contour lines $f_{-1}(\{k\})$ ($k \neq 0$) are rectangular hyperbolas; the function f does not attain maximum or minimum value on D_f

9. $D_f = \{[x, y] \in \mathbb{R}^2;\ y \neq 0\}$; f is continuous on D_f; $f_{-1}(\{0\}) = \{[x, y] \in D_f;\ x = 0\}$; $f_{-1}(\{1\}) = \{[x, y] \in D_f;\ y = x\}$; $f_{-1}(\{-1\}) = \{[x, y] \in D_f;\ y = -x\}$; the function f does not attain maximum or minimum value on D_f

10. $D_f = \{[x, y] \in \mathbb{R}^2;\ -1 \leq xy \leq 1\}$; f is continuous on D_f; $f_{-1}(\{0\}) = \{[x, y] \in \mathbb{R}^2;\ x = 0\} \cup \{[x, y] \in \mathbb{R}^2;\ y = 0\}$, $f_{-1}(\{\pi/2\}) = \{[x, y] \in \mathbb{R}^2;\ xy = 1\}$, $f_{-1}(\{-\pi/2\}) = \{[x, y] \in \mathbb{R}^2;\ xy = -1\}$, $f_{-1}(\{\pi/6\}) = \{[x, y] \in \mathbb{R}^2;\ xy = 1/2\}$; the function f attains a maximum value on D_f at points $[x, y] \in f_{-1}(\{\pi/2\})$, minimum on D_f at points $[x, y] \in f_{-1}(\{-\pi/2\})$

11. $D_f = \mathbb{R}^2$; f is continuous on $\mathbb{R}^2 \setminus \{[0, 0]\}$ and is not continuous at the point $[0, 0]$; $f_{-1}(\{0\}) = \{[x, y] \in \mathbb{R}^2;\ x = 0\} \cup \{[x, y] \in \mathbb{R}^2;\ y = 0\}$, $f_{-1}(\{1\}) = \{[x, y] \in \mathbb{R}^2;\ y = x^2\} \setminus \{[0, 0]\}$, $f_{-1}(\{-1\}) = \{[x, y] \in \mathbb{R}^2;\ y = -x^2\} \setminus \{[0, 0]\}$, $f_{-1}(\{1/2\}) = (\{[x, y] \in \mathbb{R}^2;\ y = (2+\sqrt{3})x^2\} \cup \{[x, y] \in \mathbb{R}^2;\ y = (2-\sqrt{3})x^2\}) \setminus \{[0, 0]\}$; the function f attains a maximum value on $\mathbb{R}^2$ at points $[x, y] \in f_{-1}(\{1\})$, minimum on $\mathbb{R}^2$ at points $[x, y] \in f_{-1}(\{-1\})$

12. $D_f = \mathbb{R}^2$;

$$\frac{\partial f}{\partial x}(x, y) = \frac{x}{\sqrt{x^2 + 4y^2}}, \quad \frac{\partial f}{\partial y}(x, y) = \frac{4y}{\sqrt{x^2 + 4y^2}}, \quad [x, y] \in \mathbb{R}^2 \setminus \{[0, 0]\};$$

partial derivatives at points $[0, 0]$ does not exist

13.

$$\frac{\partial f}{\partial x}(x, y) = \frac{y^3}{\sqrt{(x^2 + y^2)^3}}, \ \frac{\partial f}{\partial y}(x, y) = \frac{x^3}{\sqrt{(x^2 + y^2)^3}}, \ [x, y] \in \mathbb{R}^2 \setminus \{[0, 0]\};$$

$\frac{\partial f}{\partial x}(0, 0) = 0$, $\frac{\partial f}{\partial y}(0, 0) = 0$

14. $D_f = \{[x,y,z] \in \mathbb{R}^3;\ x>0, y>0\} \cup \{[x,y,z] \in \mathbb{R}^3;\ x<0, y<0\}$;

$$\frac{\partial f}{\partial x}(x,y,z) = \left(\frac{x}{y}\right)^z \frac{z}{x},\ \frac{\partial f}{\partial y}(x,y,z) = \left(\frac{x}{y}\right)^z \left(-\frac{z}{y}\right),$$

$$\frac{\partial f}{\partial z}(x,y,z) = \left(\frac{x}{y}\right)^z \log\frac{x}{y},$$

$$\frac{\partial^2 f}{\partial x^2}(x,y,z) = \left(\frac{x}{y}\right)^z \frac{z}{x^2}(z-1),$$

$$\frac{\partial^2 f}{\partial x \partial y}(x,y,z) = \frac{\partial^2 f}{\partial y \partial x}(x,y,z) = -\left(\frac{x}{y}\right)^z \frac{z^2}{xy},$$

$$\frac{\partial^2 f}{\partial y^2}(x,y,z) = \left(\frac{x}{y}\right)^z \frac{z}{y^2}(z+1),\quad \frac{\partial^2 f}{\partial z^2}(x,y,z) = \left(\frac{x}{y}\right)^z \left(\log\frac{x}{y}\right)^2,$$

$$\frac{\partial^2 f}{\partial y \partial z}(x,y,z) = \frac{\partial^2 f}{\partial z \partial y}(x,y,z) = -\left(\frac{x}{y}\right)^z \frac{1}{y}(z\log\frac{x}{y}+1),$$

$$\frac{\partial^2 f}{\partial x \partial z}(x,y,z) = \frac{\partial^2 f}{\partial z \partial x}(x,y,z) = \left(\frac{x}{y}\right)^z \frac{1}{x}(z\log\frac{x}{y}+1),\quad [x,y,z] \in D_f;$$

$T(x,y,z) = e^2 + 2e(x-e) - 2e^2(y-1) + e^2(z-2)$

15. $D_f = \{[x,y] \in \mathbb{R}^2;\ y \neq -x\}$;

$$\frac{\partial f}{\partial x}(x,y) = \frac{y}{x^2+y^2},\quad \frac{\partial f}{\partial y}(x,y) = \frac{-x}{x^2+y^2},$$

$$\frac{\partial^2 f}{\partial x^2}(x,y) = \frac{-2xy}{(x^2+y^2)^2},\quad \frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = \frac{x^2-y^2}{(x^2+y^2)^2},$$

$$\frac{\partial^2 f}{\partial y^2}(x,y) = \frac{2xy}{(x^2+y^2)^2},\quad [x,y] \in D_f;$$

the tangent plane to the graph of the function f exists at all points $[x,y,f(x,y)]$, where $[x,y] \in D_f$ and at the point $[1,1,f(1,1)]$ is defined by equation $2z-x+y = 0$

16. The set M is not bounded; the continuous function f thus could, but does not have to attain,its maximum and minimum values on M. But it is not difficult to realize, that the function f is continuous on M and equal to zero on both axis x, y. Next we can easily see, that on rectangular hyperbola $xy = k$, $k > 0$, is $f(x,y) = ke^{-k}$ and $\lim\limits_{k\to\infty} ke^{-k} = 0$. Hence we can derive, that f attains its maximum and minimum values on M. Suspicious points: $[x,0]$, where $x \in [0,+\infty)$; $[0,y]$, where $y \in [0,+\infty)$; $[x,1/x]$, where $x > 0$; the function f attains its maximum value on M at the points $[x,1/x]$, $x > 0$ and attains its minimum value on M at the points $[x,0]$, $x \geq 0$, and $[0,y]$, $y \geq 0$.

17. suspicious points: $[-1/\sqrt{2}, 1/\sqrt{2}]$, $[-1, 0]$, $[1, 0]$; at the point $[1, 0]$ the function f attains its maximum value on M; at the point $[-1/\sqrt{2}, 1/\sqrt{2}]$ the function f attains its minimum value on M; at the point $[-1, 0]$ there is no extrem.

18. the function f is continuous on the set$M$, which is compact. Thus the function f attains its maximum and minimum values on M. Points suspicious of being an extrem of the function are: $[1/3, 1/3]$ and all points from $H(M)$; $\max\limits_{M} f = f(1/3, 1/3) = (1/3)^3$, $\min\limits_{M} f = 0$, it is attained at all points of the boundary $H(M)$.

19. The function f is continuous on the set M, which is compact. thus the function f has a maximum and a minimum on M. suspicious points: $[\pi/2, \pi]$, $[\pi, \pi/2]$, $[\pi/2, 0]$, $[3\pi/2, 0]$, $[0, y]$ (where $y \in [0, 2\pi]$), $[\pi/4, 7\pi/4]$, $[3\pi/4, 5\pi/4]$, $[5\pi/4, 3\pi/4]$, $[7\pi/4, \pi/4]$ and $[2\pi, 0]$; maxima and minima: $\max\limits_{M} f = f(\pi/2, 0) = 1$, $\min\limits_{M} f = f(3\pi/2, 0) = f(\pi/2, \pi) = -1$.

20. the function f is continuous on the set M, which is compact. Thus the function f has a maximum and a minimum on M. Suspicious points are: $[1, 1/2, 2]$, $[0, 1/2, 2]$, $[1, 0, 2]$, $[1, 1/2, 0]$, $[11/9, 5/9, 20/9]$, $[0, 0, 2]$, $[1, 0, 0]$, $[0, 1/2, 0]$, $[0, 4/5, 16/5]$, $[3/2, 0, 5/2]$, $[3, 1, 0]$, $[0, 0, 0]$, $[4, 0, 0]$, $[0, 4, 0]$, $[0, 0, 4]$; $\max\limits_{M} f = f(0, 4, 0) = 54$, $\min\limits_{M} f = f(1, 1/2, 2) = 0$.

21. The tangent at the point $[e - 1, e]$ is described by the function $T(x) = \frac{e}{e-1}x$.

22. $y'(1) = 0$, $y''(1) = -1/3$; the function y is concave on a neighbourhood of the point 1 konkávní **23.** $y'(1) = -e$, $z'(1) = 1$ **24.** $T(x, y) = 1 + \frac{7}{5}(y + 2)$

25. $T(x, y) = 6(x - 1) - \frac{1}{2}(y + 6)$ **26.** $T(x, y) = \pi/2 - x$

27. suspicious points: $[0, 1/2]$, $[1, 3]$, $[-1, 3]$, $[3, 7]$, $[3, -7]$, $[-3, 7]$, $[-3, -7]$. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; $\max\limits_{M} f = f(3, -7) = 1080$, $\min\limits_{M} f = f(-3, -7) = -1080$.

28. Suspicious points: $[-1, 0]$, $[1, 2]$, $[1, -2]$, $[-2, 1]$, $[-2, -1]$. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; $\max\limits_{M} f = f(1, 2) = 32$, $\min\limits_{M} f = f(1, -2) = -32$.

29. Suspicious points: $[3/\sqrt{2}, \sqrt{2}]$, $[-3/\sqrt{2}, -\sqrt{2}]$, $[0, -2]$, $[-3, 0]$. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; $\max\limits_{M} f = f(3/\sqrt{2}, \sqrt{2}) = 3 + 6\sqrt{2}$, $\min\limits_{M} f = f(0, -2) = f(-3, 0) = -6$.

30. Suspicious points: $[-3/\sqrt{2}, -\sqrt{2}]$, $[0, -2]$, $[-3, 0]$, $[1/2, -1/4]$, $[12/5, -6/5]$; $[-12/5, 6/5]$. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; $\max\limits_{M} f = f(1/2, -1/4) = 1/8$, $\min\limits_{M} f = f(0, -2) = f(-3, 0) = -6$.

31. Suspicious points: $[0, 1/2]$, $[0, 1]$, $[\sqrt{15}/4, 1/4]$, $[-\sqrt{15}/4, 1/4]$, $[0, 0]$, $[1, 0]$, $[-1, 0]$. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; $\max\limits_{M} f = f(\sqrt{15}/4, 1/4) = f(-\sqrt{15}/4, 1/4) = \exp(9/8)$, $\min\limits_{M} f = f(0, 0) = f(0, 1) = 1$.

32. The set M is an intersection of a conical surface and a plane, it can be shown, that in this case it is an ellipse. The set M is compact and f is continuous on it, thus f has a maximum and a minimum on M; suspicious points: $[-3, 0, -3]$, $[1, 0, -1]$; $\max\limits_{M} f = f(-3, 0, -3) = 18$, $\min\limits_{M} f = f(1, 0, -1) = 2$.

CHAPTER 2

Matrix algebra

In this chapter we will be concerned with topics which belong to linear algebra. We will primarily deal with basic matrix operations, theory of determinants and solving systems of linear equations. All of this is very useful not only in other parts of mathematics (see for example the general formulation of the Implicit functions theorem (Theorem 1.41)), but also in varied applications in economics.

2.1. Basic operations with matrices

Definition. We call a table

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \tag{1}$$

where $a_{ij} \in \mathbb{R}$, $i = 1, \dots, m$, $j = 1, \dots, n$, the $\boldsymbol{m \times n}$ **matrix**. If $m = n$, then we call it the **square matrix of order** $\boldsymbol{n}$. The set of all $m \times n$ matrices is denoted by $M(m \times n)$.

We call an n-tuple of numbers

$$(a_{i1}, a_{i2}, \dots, a_{in}),$$

where $i \in \{1, \dots, m\}$, the $\boldsymbol{i}$**-th row** of a matrix (1) and an m-tuple of numbers

$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix},$$

where $j \in \{1, \dots, n\}$, the $\boldsymbol{j}$**-th column** of the matrix (1). The matrix (1) is also denoted by the symbol $(a_{ij})_{\substack{i=1..m \\ j=1..n}}$.

Remarks. 1. As we mentioned in the Chapter 1, we call elements of the space $\mathbb{R}^n$ also vectors. If we have a vector $\vec{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$, we can look at it as at a **row vector** , i.e. a $1 \times n$ matrix of the form

$$\begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix}$$

or as at a **column vector**, i.e. an $n \times 1$ matrix of the form

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

2. If $\vec{u}^1, \ldots, \vec{u}^m \in \mathbb{R}^n$ are row vectors, then an $m \times n$ matrix, for which the i-th row is equal to $\vec{u}^i$, $i = 1, \ldots, m$, is denoted by the symbol

$$\begin{pmatrix} \vec{u}^1 \\ \vdots \\ \vec{u}^m \end{pmatrix}.$$

Similarly, if $\vec{v}^1, \ldots, \vec{v}^n \in \mathbb{R}^m$ are column vectors, then an $m \times n$ matrix, for which j-th column is equal to $\vec{v}^j$, $j = 1, \ldots, n$, is denoted by the symbol

$$\left(\vec{v}^1, \ldots, \vec{v}^n\right).$$

Definition. We say that matrices $A = (a_{ij})_{\substack{i=1..m \\ j=1..n}}$, $B = (b_{uv})_{\substack{u=1..p \\ v=1..s}}$ are equal provided that $m = p, n = s$ and $a_{ij} = b_{ij}$ for each $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$ (i.e. matrices are of the same size and elements with the same indices are equal).

We now define two basic operations with matrices.

Definition. Let $A, B \in M(m \times n)$, $A = (a_{ij})_{\substack{i=1..m \\ j=1..n}}$, $B = (b_{ij})_{\substack{i=1..m \\ j=1..n}}$, $\lambda \in \mathbb{R}$. Then we call the following matrix the **sum of matrices** A and B

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \ldots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \ldots & a_{2n} + b_{2n} \\ a_{31} + b_{31} & \ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \ldots & a_{mn} + b_{mn} \end{pmatrix},$$

and the following matrix the **real number multiplication λ of the matrix** A

$$\lambda A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \ldots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \ldots & \lambda a_{2n} \\ \lambda a_{31} & \ldots\ldots\ldots\ldots\ldots & & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \\ \lambda a_{m1} & \lambda a_{m2} & \ldots & \lambda a_{mn} \end{pmatrix}.$$

In the following theorem there are summarized basic properties of these operations. Proofs of the assertions are very simple and we omit them.

Theorem 1. The following assertions hold:

- $\forall A, B \in M(m \times n)\colon A + B = B + A$ (commutativity),
- $\forall A, B, C \in M(m \times n)\colon A + (B + C) = (A + B) + C$ (associativity),
- there exists exactly one matrix $O \in M(m \times n)$ satisfying $O + A = A$ for each $A \in M(m \times n)$ (existence of a zero element),
- $\forall A \in M(m \times n)\ \exists C_A \in M(m \times n)\colon A + C_A = O$ (existence of an inverse element),
- $\forall A \in M(m \times n)\ \forall \lambda, \mu \in \mathbb{R}\colon (\lambda + \mu)A = \lambda A + \mu A$,
- $\forall A, B \in M(m \times n)\ \forall \lambda \in \mathbb{R}\colon \lambda(A + B) = \lambda A + \lambda B$,
- $\forall A \in M(m \times n)\ \forall \lambda, \mu \in \mathbb{R}\colon (\lambda\mu)A = \lambda(\mu A)$,
- $\forall A \in M(m \times n)\colon 1 \cdot A = A$.

Remark. It is obvious that each entry of the matrix O from the third assertion is equal to 0. We call such a matrix the **zero matrix**. It is also easy to realize that the matrix C_A from the fourth assertion is uniquely determined and is equal to $(-1)\cdot A$. Usually, we denote it by $-A$.

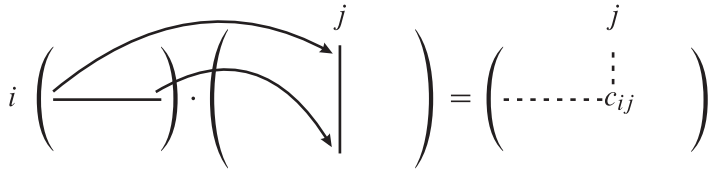
Let an important notion of matrix multiplication be defined.

Definition. Let $A = (a_{is})_{\substack{i=1..m\\ s=1..n}}$ be a $m \times n$ matrix and $B = (b_{sj})_{\substack{s=1..n\\ j=1..k}}$ be a $n \times k$ matrix. We say that a $m \times k$ matrix $A \cdot B = (c_{ij})_{\substack{i=1..m\\ j=1..k}}$ where

$$c_{ij} = \sum_{s=1}^{n} a_{is} b_{sj}, \quad i = 1, \ldots, m,\ j = 1, \ldots, k,$$

is a **matrix product** of A with B. Usually, we will write only AB instead of $A \cdot B$.

Remark. Let A be an $m \times n$ matrix and B be an $n \times k$ matrix. We calculate the entry with indices ij of the matrix AB such that we "put" the i-th row of the matrix A on j-th column of the matrix B, multiplicate the corresponding entries and add the resulting numbers together:



Let the row vectors of the matrix A be denoted by $\vec{u}^1, \ldots, \vec{u}^m$ and the column vectors of the matrix B by $\vec{v}^1, \ldots, \vec{v}^k$. The vector $\vec{u}^i$ is in fact a $1 \times n$ matrix and the vector $\vec{v}^j$ is an $n \times 1$ matrix. A matrix product of these matrices in this order is

a 1×1 matrix, whose the only entry is just a number c_{ij}. Then it is not difficult to realize that for a matrix AB the following conditions hold:

$$A \cdot B = A \cdot \underbrace{(\vec{v}^1, \ldots, \vec{v}^k)}_{B} = (A \cdot \vec{v}^1, \ldots, A \cdot \vec{v}^k),$$

$$A \cdot B = \underbrace{\begin{pmatrix} \vec{u}^1 \\ \vdots \\ \vec{u}^m \end{pmatrix}}_{A} \cdot B = \begin{pmatrix} \vec{u}^1 \cdot B \\ \vdots \\ \vec{u}^m \cdot B \end{pmatrix}.$$

The foregoing definition of a matrix multiplication may seem to be somewhat complicated. But we will see later (for example in section 2.5), that it is very natural and useful.

Example 2. Calculate the matrix product AB, where

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

Solution. The matrix A is a 2×3 matrix and the matrix B is a 3×4 matrix, thus multiplication is possible and the matrix product AB is a 2×4 matrix. According to the definition we have

$$AB = \begin{pmatrix} 1{\cdot}2{+}0{\cdot}0{+}3{\cdot}1, & 1{\cdot}1{+}0{\cdot}(-1){+}3{\cdot}0, & 1{\cdot}1{+}0{\cdot}3{+}3{\cdot}4, & 1{\cdot}0{+}0{\cdot}1{+}3{\cdot}5 \\ 2{\cdot}2{+}1{\cdot}0{+}(-1){\cdot}1, & 2{\cdot}1{+}1{\cdot}(-1){+}(-1){\cdot}0, & 2{\cdot}1{+}1{\cdot}3{+}(-1){\cdot}4, & 2{\cdot}0{+}1{\cdot}1{+}(-1){\cdot}5 \end{pmatrix} =$$

$$= \begin{pmatrix} 5 & 1 & 13 & 15 \\ 3 & 1 & 1 & -4 \end{pmatrix}.$$

♣

Theorem 3 (properties of matrix multiplication)**.** The following assertions hold:

(i) $\forall A \in M(m \times n)\ \forall B \in M(n \times k)\ \forall C \in M(k \times p)\colon (AB)C = A(BC)$ (associativity of multiplication),

(ii) $\forall A \in M(m \times n)\ \forall B, C \in M(n \times k)\colon A(B + C) = AB + AC$ (left distributivity),

(iii) $\forall A, B \in M(m \times n)\ \forall C \in M(n \times k)\colon (A + B)C = AC + BC$ (right distributivity).

(iv) There exists exactly one matrix $I \in M(n \times n)$ such that for each matrix $A \in M(n \times n)$, $IA = AI = A$ holds (existence and uniqueness of the **identity matrix** I). Above that, for the matrix I it holds:
- $\forall B \in M(m \times n)\colon BI = B$,
- $\forall C \in M(n \times k)\colon IC = C$.

Proof. (i) From the matrix multiplication definition it follows that AB is an $m \times k$ matrix and then the matrix product $(AB)C$ is defined and a result is an $m \times p$ matrix. Similarly it can be seen that BC is an $n \times p$ matrix and hence the matrix product $A(BC)$ is an $m \times p$ matrix. Both sides of the equality are thus matrices of the same size. We shall now prove that they have the same entries.

Let $A = (a_{ij})_{\substack{i=1..m \\ j=1..n}}$, $B = (b_{ij})_{\substack{i=1..n \\ j=1..k}}$, $C = (c_{ij})_{\substack{i=1..k \\ j=1..p}}$. An entry with indices ij of the matrix $(AB)C$ is equal to

$$\sum_{r=1}^{k}\left(\sum_{s=1}^{n} a_{is}b_{sr}\right)c_{rj} = \sum_{r=1}^{k}\left(\sum_{s=1}^{n} a_{is}b_{sr}c_{rj}\right)$$

and an entry with indices ij of the matrix $A(BC)$ is equal to

$$\sum_{s=1}^{n} a_{is}\left(\sum_{r=1}^{k} b_{sr}c_{rj}\right) = \sum_{s=1}^{n}\left(\sum_{r=1}^{k} a_{is}b_{sr}c_{rj}\right).$$

Adding and multiplying of real numbers are commutative and associative and hence it follows

$$\sum_{r=1}^{k}\left(\sum_{s=1}^{n} a_{is}b_{sr}c_{rj}\right) = \sum_{s=1}^{n}\left(\sum_{r=1}^{k} a_{is}b_{sr}c_{rj}\right).$$

(ii) Obviously, on both sides of the equality are $m \times k$ matrices. If $A = (a_{ij})_{\substack{i=1..m \\ j=1..n}}$, $B = (b_{ij})_{\substack{i=1..n \\ j=1..k}}$, $C = (c_{ij})_{\substack{i=1..n \\ j=1..k}}$, then an entry with indices ij of the matrix $A(B + C)$ is equal to

$$\sum_{s=1}^{n} a_{is}(b_{sj} + c_{sj}) = \sum_{s=1}^{n} a_{is}b_{sj} + \sum_{s=1}^{n} a_{is}c_{sj}.$$

Let us note that the expression $\sum_{s=1}^{n} a_{is}b_{sj}$ is equal to the entry with indices ij of the matrix AB and the expression $\sum_{s=1}^{n} a_{is}c_{sj}$ is equal to the entry with indices ij of the matrix AC. This is what had to be proved.

(iii) It can be proved similarly to the proof of (ii).

(iv) Let $I = (a_{ij})_{\substack{i=1..n \\ j=1..n}}$, where

$$a_{ij} = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j. \end{cases}$$

The matrix I is thus of the form

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

It can be easily checked, that the matrix I satisfies all required equalities.

We shall now prove the uniqueness. Let us suppose that the matrix $J \in M(n \times n)$ satisfies $AJ = JA = A$ for each matrix $A \in M(n \times n)$. However, then we get $I = IJ = J$. ∎

Remarks. 1. Let us point out that the matrix multiplication *is not* commutative. For example, if $A \in M(2 \times 3)$ and $B \in M(3 \times 4)$, then the matrix product AB is defined, but the matrix product BA is not. But the matrix multiplication is neither commutative in square matrix multiplication case:

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

2. In this text we will often appeal to the notion of an identity matrix. From the context it will be clear what size it has.

The latter matrix operation is described in the following definition.

Definition. The **transpose** of a matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \dots & a_{2n} \\ a_{31} & \dots & & & & \\ \dots & & & & & \\ a_{m1} & a_{m2} & \dots & & & a_{mn} \end{pmatrix}$$

is the matrix

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ a_{13} & a_{23} & \dots & a_{m3} \\ a_{14} & a_{24} & \dots & a_{m4} \\ a_{15} & \dots & & \\ \dots & & & \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix},$$

that is, if $A = (a_{ij})_{\substack{i=1..m \\ j=1..n}}$, then $A^T = (b_{uv})_{\substack{u=1..n \\ v=1..m}}$, where $b_{uv} = a_{vu}$ for each $u \in \{1, \dots, n\}$, $v \in \{1, \dots, m\}$.

Let us look how the transposition of the matrices relates to the foregoing operations.

Theorem 4 (properties of transposes)**.** The following assertions hold:
(i) $\forall A \in M(m \times n)\colon (A^T)^T = A$,
(ii) $\forall A, B \in M(m \times n)\colon (A + B)^T = A^T + B^T$,
(iii) $\forall A \in M(m \times k)\, \forall B \in M(k \times n)\colon (AB)^T = B^T A^T$.

Proof. The assertions (i) and (ii) are obvious.

(iii) The matrix AB is an $m \times n$ matrix, the transpose $(AB)^T$ is thus an $n \times m$ matrix. The matrix A^T is a $k \times m$ matrix, the matrix B^T is an $n \times k$ matrix, their product $B^T A^T$ is thus an $n \times m$ matrix. We obtain that on both sides of equality are matrices of the same size. We shall now prove that they have the same entries.

Let

$$A = (a_{js})_{\substack{j=1..m \\ s=1..k}}, \qquad B = (b_{si})_{\substack{s=1..k \\ i=1..n}},$$
$$A^T = (c_{pq})_{\substack{p=1..k \\ q=1..m}}, \quad B^T = (d_{rp})_{\substack{r=1..n \\ p=1..k}}.$$

Then the entry with indices ij of the matrix $(AB)^T$ is equal to the entry with indices ji of the matrix AB, i.e. it is equal to $\sum_{s=1}^{k} a_{js} b_{si}$. The entry with indices ij of the matrix $B^T A^T$ is equal to

$$\sum_{p=1}^{k} d_{ip} c_{pj} = \sum_{p=1}^{k} b_{pi} a_{jp} = \sum_{p=1}^{k} a_{jp} b_{pi},$$

which completes the proof. ∎

2.2. Invertibility and rank of a matrix

If we have a non-zero real number a, then exactly one real number b can be found such that $ab = ba = 1$ holds. We use this property of real numbers for example in solving this type of equation: $ax = c$. If we multiply both sides of equation by the number b, then we get $x = bc$. It is thus natural to ask if we can find for a given non-zero matrix $A \in M(n \times n)$ a matrix $B \in M(n \times n)$ satisfying $AB = BA = I$. Generally, the answer is negative. It is not difficult to make sure that for the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

we can not find a matrix $B \in M(2 \times 2)$ satisfying $AB = BA = I$. It is therefore useful to separate the matrices for which the answer to the given question is positive.

Definition. Let $A \in M(n \times n)$. We say that a matrix A is **invertible** if there exists $B \in M(n \times n)$ such that

$$AB = BA = I. \tag{2}$$

Remarks. 1. But are there any invertible matrices? It can be directly seen that for example an identity matrix of any size is invertible. In this section we will prove relatively simple criterion, which determine if a matrix is invertible or not.

2. Let us note that if $AB = I$ holds for $A, B \in M(n \times n)$, then also $BA = I$ holds. We prove this assertion in the last section of this chapter.

Let us note that if a matrix A is invertible, then there exists exactly one matrix B satisfying $AB = BA = I$. Since if we have an invertible matrix $A \in M(n \times n)$ and matrices $B, \tilde{B} \in M(n \times n)$, which satisfy $AB = BA = I$, $A\tilde{B} = \tilde{B}A = I$, then $B = IB = (\tilde{B}A)B = \tilde{B}(AB) = \tilde{B}I = \tilde{B}$. This observation shows that the following definition and notation are correct.

Definition. Let A be an invertible matrix. We call the matrix B satisfying (2) the **inverse** of a matrix A. We denote it by A^{-1}.

Theorem 5. Let $A, B \in M(n \times n)$ be invertible matrices. Then:
(i) A^{-1} is an invertible matrix and $(A^{-1})^{-1} = A$,
(ii) A^T is an invertible matrix and $(A^T)^{-1} = (A^{-1})^T$,
(iii) AB is an invertible matrix and $(AB)^{-1} = B^{-1}A^{-1}$.

Proof. The assertion (i) is obvious.

(ii) $AA^{-1} = A^{-1}A = I$ holds and hence $(AA^{-1})^T = (A^{-1}A)^T = I^T$ and from Theorem 4 it follows $(A^{-1})^T A^T = A^T(A^{-1})^T = I$. Hence, our assertion follows.

(iii) $AA^{-1} = A^{-1}A = I$ and $BB^{-1} = B^{-1}B = I$ hold. We thus have

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I,$$
$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I.$$

Let us realize that we use the associativity of a matrix multiplication. From the previous relations we obtain the assertion. ∎

Definition. Let $k, n \in \mathbb{N}$ and $\vec{v}^1, \ldots, \vec{v}^k \in \mathbb{R}^n$. We say that a vector $\vec{u} \in \mathbb{R}^n$ **is the linear combination of vectors $\vec{v}^1, \ldots, \vec{v}^k$ with coefficients $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$** provided that

$$\vec{u} = \lambda_1 \vec{v}^1 + \cdots + \lambda_k \vec{v}^k.$$

In this case we also say that the **linear combination of vectors $\vec{v}^1, \ldots, \vec{v}^k$ with coefficients $\lambda_1, \ldots, \lambda_k$ is equal to $\vec{u}$**.

If $\lambda_1 = \cdots = \lambda_k = 0$, then we call it the **trivial linear combination** of the vectors $\vec{v}^1, \ldots, \vec{v}^k$; if any of the coefficient is non-zero then it is the **non-trivial linear combination**.

Definition. We say that the vectors $\vec{v}^1, \ldots, \vec{v}^k \in \mathbb{R}^n$ are **linearly dependent** if there exists any netrivial linear combination which is equal to a zero vector. We say that the vectors $\vec{v}^1, \ldots, \vec{v}^k \in \mathbb{R}^n$ are **linearly independent**, if they are not linearly dependent, i.e. if

$$\forall \lambda_1, \ldots, \lambda_k \in \mathbb{R}\colon \lambda_1 \vec{v}^1 + \cdots + \lambda_k \vec{v}^k = \vec{o} \Rightarrow \lambda_1 = \lambda_2 = \cdots = \lambda_k = 0.$$

In other words among all linear combinations of vectors $\vec{v}^1, \ldots, \vec{v}^k$ only the trivial linear combinations is equal to a zero vector.

Remark. Vectors $\vec{v}^1, \ldots, \vec{v}^k$ are linearly dependent if and only if one of them is a linear combination of the others.

Since if $\lambda_1 \vec{v}^1 + \cdots + \lambda_k \vec{v}^k$ is a netrivial linear combination which equals to a zero vector, then for a certain i we have $\lambda_i \neq 0$ and we could write

$$\vec{v}^i = -\frac{\lambda_1}{\lambda_i}\vec{v}^1 - \cdots - \frac{\lambda_{i-1}}{\lambda_i}\vec{v}^{i-1} - \frac{\lambda_{i+1}}{\lambda_i}\vec{v}^{i+1} - \cdots - \frac{\lambda_k}{\lambda_i}\vec{v}^k.$$

We shall now prove the converse implication. If a certain vector $\vec{v}^i$ is a linear combination of the other vectors, i.e.

$$\vec{v}^i = \alpha_1 \vec{v}^1 + \cdots + \alpha_{i-1}\vec{v}^{i-1} + \alpha_{i+1}\vec{v}^{i+1} + \cdots + \alpha_k \vec{v}^k,$$

then by adding a vector $(-1) \cdot \vec{v}^i$ to both sides of equality we get a netrivial linear combination which is equal to a zero vector.

Definition. Let $A \in M(m \times n)$. The **rank of the matrix** A is the maximal number of linearly independent rows, i.e. the rank is equal to $k \in \mathbb{N}$ if

(i) there exist k linearly independent row vectors of the matrix A and
(ii) every l-tuple of row vectors of the matrix A, where $l > k$, is linearly dependent.

The rank of a zero matrix is equal to zero. The rank of a matrix A is denoted by $h(A)$.

Remark. It is obvious that a non-zero vector is linearly independent and if in an l-tuple of vectors there are at least two of them the same, then this l-tuple is linearly dependent. Hence, the rank is correctly defined for each matrix and is at most equal to its number of rows.

Most matrices has not obvious rank at the first sight. But for some matrices, it is very easy to determine it.

Definition. We say that $A \in M(m \times n)$ is in the **row echelon form**, if for each $i \in \{2, \ldots, m\}$ it holds that the i-th row of the matrix A is a zero vector or starts with more zeros than $(i-1)$-th row.

It is not difficult to realize that the rank of a matrix in a row echelon form is equal to the number of non-zero rows. We show a method how to use this observation to determine the rank of a general matrix.

Definition. The **elementary row operations** of a matrix A are:

(i) switching any two rows,
(ii) multiplying a row by a non-zero number,
(iii) adding multiple of one row to another one.

Definition. We call the **transformation** a finite sequence of elementary row operations. If a matrix $B \in M(m \times n)$ originated from a matrix $A \in M(m \times n)$ applying a transformation T on a matrix A, then we use this notation $A \overset{T}{\rightsquigarrow} B$.

The assertions (i) and (iii) in the following theorem show how to determine the rank of a general matrix. We could transform a given matrix to a matrix in a row echelon form (assertion (i)) whose rank is then equal to the rank of the original matrix (assertion (iii)).

Theorem 6 (properties of transformation)**.**

(i) Let $A \in M(m \times n)$. Then there exists a transformation T which alter the matrix A to a matrix in a row echelon form.
(ii) Let T_1 be a transformation applicable on $m \times n$ matrices. Then there exists a transformation T_2 applicable on $m \times n$ matrices such that for each two matrices $A, B \in M(m \times n)$ $A \overset{T_1}{\rightsquigarrow} B$ holds if and only if $B \overset{T_2}{\rightsquigarrow} A$.
(iii) Let $A, B \in M(m \times n)$ and matrix A became matrix B by a transformation. Then $h(A) = h(B)$.

Proof. (i) We prove the result by applying mathematical induction on m. If $m = 1$, then we does not have to prove anything, since the matrix has only one row and thus it is in a row echelon form. Let us suppose that the assertion holds for all matrices with m rows. Let $A \in M((m+1) \times n)$. If A is a zero matrix, then it is also in a row echelon form. Let thus A be a non-zero matrix. Let us find the smallest $j \in \{1, \ldots, n\}$ such that the j-th column of the matrix A contains non-zero entry. Let this entry be in an i-th row. Then we swap the i-th row with the first. Let the newly originated matrix be denoted by B and its entries by b_{ij}. Let us take $s \in \{2, \ldots, m+1\}$ and add a $(-b_{sj}/b_{1j})$-multiple of the first row to the s-th row of the matrix B. The originated matrix has zero in the position sj. Let us repeat this process for each $s \in \{2, \ldots, m+1\}$. This way we get a matrix C which has zero entries in the first $j-1$ columns and in j-th column there is a only one entry, just on the first position. From the induction assumption it follows that there exists a transformation of only the second to the $(m+1)$-st row, which alter the matrix C to be in a row echelon form.

(ii) For the purpose of this proof we use the following notation. If T is a row elementary operation of swapping an i-th and a j-th row, then take T^{-1} equal to T. If T is a row elementary operation of multiplying an i-th row by a non-zero number λ, then T^{-1} stands for an elementary operation of multiplying the i-th row by a (non-zero) number $1/\lambda$. Finally, if T is a row elementary operation of adding a λ-multiple of a j-th row to an i-th row and $i \neq j$, then T^{-1} stands for an elementary operation of adding $(-\lambda)$-multiple of the j-th row to the i-th row.

Let us suppose first that the transformation T_1 consists of one row elementary operation; then it is obvious that the transformation $T_2 = T_1^{-1}$ has the required property. Now if T_1 is a transformation consisting of a sequence of row elementary operations $P_1, P_2, \ldots, P_k$, then let us denote a transformation consisting of a sequence of $P_k^{-1}, \ldots, P_2^{-1}, P_1^{-1}$ by T_2. Then for matrices $A, B \in M(m \times n)$ follows

$$A \overset{P_1}{\rightsquigarrow} A_1 \overset{P_2}{\rightsquigarrow} \ldots \overset{P_{k-1}}{\rightsquigarrow} A_{k-1} \overset{P_k}{\rightsquigarrow} B, \quad \text{if and only if}$$
$$B \overset{P_k^{-1}}{\rightsquigarrow} A_{k-1} \overset{P_{k-1}^{-1}}{\rightsquigarrow} \ldots \overset{P_2^{-1}}{\rightsquigarrow} A_1 \overset{P_1^{-1}}{\rightsquigarrow} A,$$

and this is what had to be proved.

(iii) It could be seen that the row elementary operations of the first and the second type does not change the rank of a matrix.

We shall now prove that the row elementary operation of the third type does not reduce the rank of a matrix. Let us suppose that $h(A) = l$. Let the row vectors of the matrix A be denoted by order $\vec{v}^1, \ldots, \vec{v}^m$. We can assume without loss of generality that exactly the vectors $\vec{v}^1, \ldots, \vec{v}^l$ are linearly independent. Let A be transformed by a row elementary operation of the third type. If this operation does not change any of the vectors $\vec{v}^1, \ldots, \vec{v}^l$, then the rank of newly originated matrix cannot be smaller than l.

Let us suppose now that we added λ-multiple of some vector to some another vector of $\vec{v}^1, \ldots, \vec{v}^l$. We can assume without loss of generality that we added vector $\lambda\vec{v}^i$, $i \in \{1, \ldots, m\}$ to the l-th vector provided $i \neq l$. If the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}$, $\vec{v}^i$ are linearly independent, then the proof is completed, since the newly originated matrix has the rank at least l. Let us assume now that the given vectors are linearly dependent. From that we derive the linear independence of the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}$, $\vec{v}^l + \lambda\vec{v}^i$. Let us take any linear combination of the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}$, $\vec{v}^l + \lambda\vec{v}^i$ which equals to a zero vector:

$$\mu_1\vec{v}^1 + \cdots + \mu_{l-1}\vec{v}^{l-1} + \mu_l(\vec{v}^l + \lambda\vec{v}^i) = \vec{o}. \tag{3}$$

From the linear independence of the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}$ and from the linear dependence of the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}$, $\vec{v}^i$ follows that $\vec{v}^i$ is a linear combination

$\vec{v}^1, \ldots, \vec{v}^{l-1}$, i.e.

$$\vec{v}^i = \sum_{j=1}^{l-1} \tau_j \vec{v}^j.$$

By plugging this into the (3) and manipulating the equation, we obtain

$$(\mu_1 + \mu_l \lambda \tau_1)\vec{v}^1 + \cdots + (\mu_{l-1} + \mu_l \lambda \tau_{l-1})\vec{v}^{l-1} + \mu_l \vec{v}^l = \vec{o}.$$

The vectors $\vec{v}^1, \ldots, \vec{v}^l$ are linearly independent and therefore it must be

$$\mu_1 + \mu_l \lambda \tau_1 = 0, \ldots, \mu_{l-1} + \mu_l \lambda \tau_{l-1} = 0 \quad \text{and} \quad \mu_l = 0.$$

Hence, $\mu_1 = \cdots = \mu_l = 0$ and this proves the linear independence of the vectors $\vec{v}^1, \ldots, \vec{v}^{l-1}, \vec{v}^l + \lambda \vec{v}^i$.

Thus, if the matrix A becomes a matrix B by a certain transformation, then $h(A) \leq h(B)$. Acoording to the assertion (ii) we can transform B to A and therefore $h(B) \leq h(A)$. Thus, we get $h(A) = h(B)$. ∎

Remarks. 1. Let us realize that the proof of the point (i) gives us at the same time instruction how to transform a given matrix to be in a row echelon form.

2. We can define column elementary operations similarly to row elementary operations. It can be shown that after column elementary operation the linearly dependent rows of an original matrix are linearly dependent also in a new matrix and linearly independent rows stays independent. Hence, the column elementary operations does not change the rank of a matrix.

If we want to determine the rank of a matrix, we can thus use both rows and column elementary operations. But in certain situations (e.g. solving linear systems of equations, see the section 2.4) we can use only rows elementary operations.

3. Moreover, it can be shown that $h(A) = h(A^T)$ holds provided $A \in M(m \times n)$. Since $(A^T)^T = A$, it is sufficient to show that $h(A^T) \geq h(A)$. If a matrix A is in a row echelon form, then the proof of this inequality is simple. According to Theorem 6 we can transform A to a matrix B in a row echelon form and $h(A) = h(B)$. If we perform corresponding column elementary operation on A^T, we get B^T. According to the previous point $h(A^T) = h(B^T)$ holds. Due to $h(B^T) \geq h(B)$, we obtain $h(A^T) \geq h(A)$.

Example 7. Determine the rank of the matrix $h(A)$

$$A = \begin{pmatrix} 1 & 3 & -2 & 2 & 4 \\ 2 & 7 & 3 & 0 & -1 \\ -1 & 1 & 3 & 1 & 5 \\ -2 & 1 & -1 & 6 & 19 \end{pmatrix}.$$

Solution. The rank of the matrix A cannot be bigger than 4. We transform A to a matrix in a row echelon form.

1. Let us copy the first row, add a (-2)-multiple of the first row to the second one, add 1-multiple of the first row to the third one and add 2-multiple of the first row to the fourth one. This way we get the matrix

$$\begin{pmatrix} 1 & 3 & -2 & 2 & 4 \\ 0 & 1 & 7 & -4 & -9 \\ 0 & 4 & 1 & 3 & 9 \\ 0 & 7 & -5 & 10 & 27 \end{pmatrix}.$$

2. Let us copy the first and the second row, add a (-4)-multiple of the second row to the third one and add a (-7)-multiple of the second row to the fourth one. Now we get the matrix

$$\begin{pmatrix} 1 & 3 & -2 & 2 & 4 \\ 0 & 1 & 7 & -4 & -9 \\ 0 & 0 & -27 & 19 & 45 \\ 0 & 0 & -54 & 38 & 90 \end{pmatrix}.$$

3. Let us copy the first three rows and add a (-2)-multiple of the third row to the fourth one. We get the matrix

$$\begin{pmatrix} 1 & 3 & -2 & 2 & 4 \\ 0 & 1 & 7 & -4 & -9 \\ 0 & 0 & -27 & 19 & 45 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

A transformation does not change the rank of a matrix and thus $h(A) = 3$. ♣

Example 8. Determine the rank $h(A)$ of the matrix

$$A = \begin{pmatrix} 1 & 0 & 5 & -1 \\ 0 & 3 & -2 & 5 \\ 2 & 9 & 4 & a \\ 1 & 15 & b & 24 \end{pmatrix}$$

in dependance on real parameters a, b.

Solution. We use a suitable transformation and get these matrices consecutively

$$\begin{pmatrix} 1 & 0 & 5 & -1 \\ 0 & 3 & -2 & 5 \\ 0 & 9 & -6 & a+2 \\ 0 & 15 & b-5 & 25 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 5 & -1 \\ 0 & 3 & -2 & 5 \\ 0 & 0 & 0 & a-13 \\ 0 & 0 & b+5 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 5 & -1 \\ 0 & 3 & -2 & 5 \\ 0 & 0 & b+5 & 0 \\ 0 & 0 & 0 & a-13 \end{pmatrix}.$$

Now it is obvious that $h(A) = 2$ in case that $a = 13$ and $b = -5$. Next, if $a = 13$ and $b \neq -5$ or $a \neq 13$ and $b = -5$, then the rank of the matrix is $h(A) = 3$. If $a \neq 13, b \neq -5$, then $h(A) = 4$. ♣

Theorem 9 (about transformation and matrix product)**.** Let $A \in M(m \times k)$, $B \in M(k \times n)$, $C \in M(m \times n)$ and $AB = C$ hold. Let T be a transformation and $A \overset{T}{\rightsquigarrow} A'$, $C \overset{T}{\rightsquigarrow} C'$. Then $A'B = C'$.

Proof. Let us assume without loss of generality that T is only one row elementary operation. Let the rows of the matrix A be denoted by $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m$ and the rows of the matrix C by $\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_m$. According to the remark following the definition of matrix product on page 63 we have

$$\vec{c}_j = \vec{a}_j B, \quad j = 1, \ldots, m.$$

For row elementary operations of the first and the second type the situation is thus clear.

Let us prove the assertion in case that we add λ-multiple of the p-th row to the q-th row, $p \neq q$. Then the q-th row of the matrix A' is equal to $\vec{a}_q + \lambda\vec{a}_p$ and the q-th row of the matrix C' is equal to $\vec{c}_q + \lambda\vec{c}_p$. Then we have

$$(\vec{a}_q + \lambda\vec{a}_p)\, B = \vec{a}_q B + \lambda\vec{a}_p B = \vec{c}_q + \lambda\vec{c}_p.$$

Since the other rows did not change, we get $A'B = C'$ and this is what had to be proved. ∎

Lemma 10. Let $A \in M(n \times n)$ and $h(A) = n$. Then there exists a transformation T which alter A to I.

Proof. According to Theorem 6(i) we can transform A to a matrix A' (with entries a'_{ij}) in a row echelon form. From Theorem 6(iii) and the assumption $h(A) = n$ we obtain $h(A') = n$. Hence, $a'_{ii} \neq 0$, $i = 1, \ldots, n$. If we multiply the i-th row by a non-zero number $1/a'_{ii}$, $i = 1, \ldots, n$, then we get the entries on the diagonal equal to 1. Thus, we can directly suppose that $a'_{ii} = 1$, $i = 1, \ldots, n$.

Now let us take consecutively a $(-a'_{i,n})$-multiple of the n-th row and add it to the i-th row for $i = 1, \ldots, n-1$. The newly originated matrix B_1 has the columns equal to the columns of the matrix A' except for the n-th column which is equal to the n-th column of an identity matrix. In the next step let us take consecutively a $(-a'_{i,n-1})$-multiple of the $(n-1)$-st row and add it to the i-th row for $i = 1, \ldots, n-2$. The newly originated matrix B_2 has the columns equal to the columns of the matrix A' except for the two last columns which are equal to the corresponding columns of an identity matrix. We repeat this procedure and finally we get (using row elementary operations of the third type) the matrices $B_3, \ldots, B_{n-1}$. Then $B_{n-1} = I$ holds and this would complete the proof. ∎

Remark. Let us realize that this proof gives us also instruction how to find such a transformation.

The latter theorem gives us an important characterization of regular matrices.

Theorem 11 (invertibility and rank)**.** Let $A \in M(n \times n)$. The matrix A is regular if and only if $h(A) = n$.

Proof. Let us assume first that the matrix A is invertible, but $h(A) < n$. Thus, we can find an inverse A^{-1} of a matrix A. Let us find a transformation T which alter the matrix A to a matrix in a row echelon form S. Let a matrix, transformed by T from I, be denoted by B. According to Theorem 9, $SA^{-1} = B$ holds, since $AA^{-1} = I$. Since $h(A) < n$ and a transformation does not change the rank of a matrix, it follows that $h(S) < n$, and thus the last row of the matrix S is a zero vector. Therefore also the last row of a matrix B is a zero vector and hence $h(B) < n$. Simultanously, we have $h(B) = h(I) = n$ and that is a contradiction. This would complete the proof, that the rank of an invertible matrix is n.

Let now $h(A) = n$. Then according to Lemma 10 there exists a transformation T_1 which alter A to I. Let us apply T_1 on I and let the resulting matrix be denoted by B. According to Theorem 6(ii) we can find a transformation T_2 to the transformation T_1. Then $A \overset{T_1}{\rightsquigarrow} I \overset{T_2}{\rightsquigarrow} A$ and $I \overset{T_1}{\rightsquigarrow} B \overset{T_2}{\rightsquigarrow} I$ holds. By using Theorem 9 and the transformation T_2 on the equality $IB = B$ we get the equality $AB = I$. Similarly, from the equality $IA = A$ and by using the transformation T_1 we obtain an equality $BA = I$. The matrix A is thus invertible. ∎

Method of matrix inversion. The second part of the just finished proof gives us instruction for matrix inverse calculation. Let a matrix $A \in M(n \times n)$ be invertible. Let us transform the matrix A to an identity matrix I (see Lemma 10). Using the same row elementary operations simultanously on I we get a matrix B satisfying $AB = BA = I$. Thus $B = A^{-1}$ holds. In other words – if we tranform A to a I, then the same transformation alter I to A^{-1}.

Example 12. Find an inverse of the matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 0 & -1 & 0 \end{pmatrix}.$$

Solution. Usually, we proceed as described below: from matrices $A, I \in M(n \times n)$ we form a matrix $(A|I) \in M(n \times 2n)$, which we transform by suitable row elementary operations to $(I|A^{-1})$. Let us calculate:

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{array}\right),$$

let us subtract the 1st row from the 2nd row:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{array}\right),$$

let us add the 2nd row to the 3rd row:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right),$$

let us subtract the 3rd row from the 2nd row:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right),$$

let us subtract the 3rd row from the 1st row:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 2 & -1 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right).$$

The matrix

$$\begin{pmatrix} 2 & -1 & -1 \\ 0 & 0 & -1 \\ -1 & 1 & 1 \end{pmatrix}$$

is thus the searched inverse of the original matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 0 & -1 & 0 \end{pmatrix}.$$

♣

Example 13. Find an inverse of the matrix

$$A = \begin{pmatrix} -18 & -16 & -11 & 12 \\ -6 & -6 & -4 & 5 \\ -11 & -10 & -7 & 8 \\ -1 & -1 & -1 & 1 \end{pmatrix}.$$

Solution. Let us use row elementary operations on the 4×8 matrix:

$$(A|I) = \left(\begin{array}{cccc|cccc} -18 & -16 & -11 & 12 & 1 & 0 & 0 & 0 \\ -6 & -6 & -4 & 5 & 0 & 1 & 0 & 0 \\ -11 & -10 & -7 & 8 & 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 & 0 & 0 & 0 & 1 \end{array}\right).$$

In the matrix $(A|I)$, let us swap the 1st and the 4th row:

$$\left(\begin{array}{cccc|cccc} -1 & -1 & -1 & 1 & 0 & 0 & 0 & 1 \\ -6 & -6 & -4 & 5 & 0 & 1 & 0 & 0 \\ -11 & -10 & -7 & 8 & 0 & 0 & 1 & 0 \\ -18 & -16 & -11 & 12 & 1 & 0 & 0 & 0 \end{array}\right),$$

let us multiply the 1st row by -1:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ -6 & -6 & -4 & 5 & 0 & 1 & 0 & 0 \\ -11 & -10 & -7 & 8 & 0 & 0 & 1 & 0 \\ -18 & -16 & -11 & 12 & 1 & 0 & 0 & 0 \end{array}\right),$$

let us add suitable multiples consecutively to the 2nd, 3rd and 4th row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & -1 & 0 & 1 & 0 & -6 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 2 & 7 & -6 & 1 & 0 & 0 & -18 \end{array}\right),$$

let us swap the 2nd and the 3rd row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 0 & 2 & -1 & 0 & 1 & 0 & -6 \\ 0 & 2 & 7 & -6 & 1 & 0 & 0 & -18 \end{array}\right),$$

let us subtract a double of the 2nd row from the 4th row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 0 & 2 & -1 & 0 & 1 & 0 & -6 \\ 0 & 0 & -1 & 0 & 1 & 0 & -2 & 4 \end{array}\right),$$

let us swap the 3rd and the 4th row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 0 & -1 & 0 & 1 & 0 & -2 & 4 \\ 0 & 0 & 2 & -1 & 0 & 1 & 0 & -6 \end{array}\right),$$

let us add a double of the 3rd row to the 4th row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 0 & -1 & 0 & 1 & 0 & -2 & 4 \\ 0 & 0 & 0 & -1 & 2 & 1 & -4 & 2 \end{array}\right),$$

let us multiply the 3rd and the 4th row by -1:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & -3 & 0 & 0 & 1 & -11 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right),$$

let us add a triple of the 4th row to the 2nd row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 4 & 0 & -6 & -3 & 13 & -17 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right),$$

let us add 4th row to the 1st row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 0 & -2 & -1 & 4 & -3 \\ 0 & 1 & 4 & 0 & -6 & -3 & 13 & -17 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right),$$

let us subtract a quadruple of the 3rd row from the 2nd row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 0 & -2 & -1 & 4 & -3 \\ 0 & 1 & 0 & 0 & -2 & -3 & 5 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right),$$

let us subtract the 3rd row from the 1st row:

$$\left(\begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & -1 & -1 & 2 & 1 \\ 0 & 1 & 0 & 0 & -2 & -3 & 5 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right),$$

and finally let us subtract 2nd row from the 1st row:

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 2 & -3 & 2 \\ 0 & 1 & 0 & 0 & -2 & -3 & 5 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 & 2 & -4 \\ 0 & 0 & 0 & 1 & -2 & -1 & 4 & -2 \end{array}\right).$$

Thus we have

$$A^{-1} = \begin{pmatrix} 1 & 2 & -3 & 2 \\ -2 & -3 & 5 & -1 \\ -1 & 0 & 2 & -4 \\ -2 & -1 & 4 & -2 \end{pmatrix}.$$

Let us note that we can check our calculation by multiplying AA^{-1}. If we proceeded correctly, the result must be I. ♣

2.3. Determinants

Now we will study the notion of determinant, which plays an important role in mathematics.

Definition. Let $A \in M(n \times n)$. We denote by the symbol A_{ij} a $(n-1) \times (n-1)$ matrix which becomes from A by omitting the i-th row and the j-th column.

Definition. Let $A = (a_{ij})_{\substack{i=1..n \\ j=1..n}}$. We define the **determinant of the matrix** A by:

$$\det A = \begin{cases} a_{11} & \text{if } n = 1, \\ \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det A_{i1} & \text{if } n > 1. \end{cases}$$

For $\det A$, we also use the symbol

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Remarks. 1. According to the mathematical induction princip the determinant notion is defined by the previous definition for each square matrix of order $n \in \mathbb{N}$.

2. Let us calculate a determinant of the 2×2 matrix A:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a \det(d) - c \det(b) = ad - bc.$$

It can be seen that for $n = 2$ our definition is consistent with the high school formula.

The next part of the section will be dedicated to derivation of some basic properties of a determinant.

Theorem 14. Let $j, n \in \mathbb{N}$, $j \leq n$, and matrices $A, B, C \in M(n \times n)$ be the same at all rows except the j-th. Let us suppose that j-th row of the matrix A is equal to sum of the j-th row of the matrix B and the j-th row of the matrix C. Then $\det A = \det B + \det C$. We can reformulate this assertion in the following way:

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{j-1,1} & a_{j-1,2} & \dots & a_{j-1,n} \\ u_1+v_1 & u_2+v_2 & \dots & u_n+v_n \\ a_{j+1,1} & a_{j+1,2} & \dots & a_{j+1,n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{j-1,1} & a_{j-1,2} & \dots & a_{j-1,n} \\ u_1 & u_2 & \dots & u_n \\ a_{j+1,1} & a_{j+1,2} & \dots & a_{j+1,n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{j-1,1} & a_{j-1,2} & \dots & a_{j-1,n} \\ v_1 & v_2 & \dots & v_n \\ a_{j+1,1} & a_{j+1,2} & \dots & a_{j+1,n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Proof. We prove the result by applying mathematical induction on n. For $n = 1$ the assertion is obvious. Now let us assume that for $n-1$ the assertion holds and we derive its validity for n. From the definition we have[1]

$$\det A = \sum_{\substack{1\le i\le n\\ i\neq j}} (-1)^{i+1} a_{i1} \det A_{i1} + (-1)^{j+1}(u_1 + v_1) \det A_{j1}.$$

According to the induction assumption for each $i \in \{1, \dots, n\}$, $i \neq j$ the equality $\det A_{i1} = \det B_{i1} + \det C_{i1}$ holds. Above that, $A_{j1} = B_{j1} = C_{j1}$ obviously holds. Thus,

$$\begin{aligned}\det A &= \sum_{\substack{1\le i\le n\\ i\neq j}} (-1)^{i+1} a_{i1}(\det B_{i1} + \det C_{i1}) + (-1)^{j+1}(u_1 + v_1) \det A_{j1} = \\ &= \sum_{\substack{1\le i\le n\\ i\neq j}} (-1)^{i+1} a_{i1} \det B_{i1} + (-1)^{j+1} u_1 \det B_{j1} + \\ &\quad + \sum_{\substack{1\le i\le n\\ i\neq j}} (-1)^{i+1} a_{i1} \det C_{i1} + (-1)^{j+1} v_1 \det C_{j1} = \det B + \det C.\end{aligned}$$

■

Theorem 15 (determinant and transformation)**.** Let $A, A' \in M(n \times n)$.

(i) If the matrix A' becomes from the A by multiplying one row by a real number, then $\det A' = \mu \det A$.
(ii) If the matrix A' becomes from the A by swapping of two rows (in other words, we use a row elementary operation of the first type), then $\det A' = -\det A$.
(iii) If the matrix A' becomes from the A by adding μ-multiple of one row to another row (in other words, we use a row elementary operation of the third type), then $\det A' = \det A$.
(iv) If the matrix A' becomes from the A by a transformation, then $\det A \neq 0$ holds if and only if $\det A' \neq 0$.

Proof. Let us use the following notation: $A = (a_{ij})_{\substack{i=1..n\\ j=1..n}}$ and $A' = \left(a'_{ij}\right)_{\substack{i=1..n\\ j=1..n}}$.

(i) We use the mathematical induction. For $n = 1$ the assertion is obvious. Let us assume that for each $(n-1) \times (n-1)$ matrix, where $n > 1$, the assertion holds. Let the matrix A' becomes from A by multiplying the j-th row of the matrix A by

[1] We denote by the symbol $\sum_{\substack{1\le i\le n\\ i\neq j}}$ the sum over all indices $i \in \{1, \dots, n\} \setminus \{j\}$.

μ. Then:

$$\begin{aligned}\det A' &= \sum_{i=1}^{n}(-1)^{i+1}a'_{i1}\det A'_{i1} = \\ &= \sum_{\substack{1\le i\le n\\ i\neq j}}(-1)^{i+1}a_{i1}\det A'_{i1} + (-1)^{j+1}\mu a_{j1}\det A'_{j1} = \\ &= \sum_{\substack{1\le i\le n\\ i\neq j}}(-1)^{i+1}a_{i1}\mu\det A_{i1} + (-1)^{j+1}\mu a_{j1}\det A_{j1} = \mu\det A.\end{aligned}$$

In the third equality we used the induction assumption and an obvious fact that $A'_{j1} = A_{j1}$.

(ii) We prove this part of the theorem by applying mathematical induction as well. For $n = 1$ there is nothing to prove and for $n = 2$ we can verify the assertion by a direct calculation. Let us assume that the assertion was already proved for $n - 1$. Let the matrix $A' \in M(n \times n)$ becomes from A by swapping of the k-th and the l-th row, $k < l$. Let us calculate

$$\begin{aligned}\det A' = \sum_{\substack{1\le i\le n\\ i\neq k,l}}(-1)^{i+1}a_{i1}\det A'_{i1} + \\ + (-1)^{k+1}a_{l1}\det A'_{k1} + (-1)^{l+1}a_{k1}\det A'_{l1}.\end{aligned} \tag{4}$$

If we apply the induction assumption to the matrices A'_{i1}, $i \neq k, l$, we obtain $\det A'_{i1} = -\det A_{i1}$. If we swap in the matrix A'_{l1} the row k with the row $(k + 1)$, in the newly originated matrix the row $(k + 1)$ with the row $(k + 2)$ etc., then after $l - k - 1$ step we get the matrix A_{k1}. From that and from the induction assumption it follows that $\det A_{k1} = (-1)^{l-k-1}\det A'_{l1}$, in other words $\det A'_{l1} = (-1)^{-l+k+1}\det A_{k1}$ holds. Similarly we can derive that $\det A'_{k1} = (-1)^{-l+k+1}\det A_{l1}$ holds. Let us apply this in (4):

$$\begin{aligned}\det A' = -\sum_{\substack{1\le i\le n\\ i\neq k,l}}(-1)^{i+1}a_{i1}\det A_{i1} + (-1)^{k+1}(-1)^{-l+k+1}a_{l1}\det A_{l1} + \\ + (-1)^{l+1}(-1)^{-l+k+1}a_{k1}\det A_{k1} = -\det A.\end{aligned}$$

(iii) Let us suppose that the matrix A' becomes from A by adding a μ-multiple of the k-th row to the l-th row. Then we have $\det A' = \det A + \mu\det B$, where the matrix B becomes from the matrix A by replacing the l-th row by the k-th row (according to Theorem 14 and the point (i)). The matrix B has the l-th row equal to the k-th row. Thus, if we swap the k-th and the l-th rows, the matrix B does not change. However, according to the assertion (ii) $\det B = -\det B$ holds, i.e. $\det B = 0$. This would complete the proof of (iii).

(iv) Proof of this assertion follows easily from (i)–(iii). ∎

Remark. Let us note that from Theorem 15(i) it follows that a determinant of a matrix with a zero row is equal to zero. From Theorem 15(ii) follows that a determinant of a matrix with two identical rows is also equal to zero (see also the proof of the part (iii) of Theorem 15).

Next, we will show a simple theorem, which gives us (together with Theorem 15) another way, how to calculate determinants. We will start with the following definition.

Definition. Let $A = (a_{ij})_{\substack{i=1..n \\ j=1..n}} \in M(n \times n)$. We say that A is the **upper triangular matrix** provided that $a_{ij} = 0$ for $i > j$, $i, j \in \{1, \dots, n\}$. We say that A is the **lower triangular matrix** provided that $a_{ij} = 0$ for $i < j$, $i, j \in \{1, \dots, n\}$.

Theorem 16. Let $A = (a_{ij})_{\substack{i=1..n \\ j=1..n}} \in M(n \times n)$ be an upper (a lower) triangular matrix. Then

$$\det A = a_{11} \cdot a_{22} \cdot \dots \cdot a_{nn}.$$

Proof. Let us use the mathematical induction. For $n = 1$ the assertion is obvious. For $n \geq 2$, let us assume that for every upper (lower, respectively) $(n-1) \times (n-1)$ triangular matrix the assertion holds. Let $A \in M(n \times n)$ be upper (lower, respectively) triangular matrix. According to the definition of a determinant we have

$$\det A = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det A_{i1}.$$

If A is an upper triangular matrix, then $a_{i1} = 0$ for $i = 2, \dots, n$, and thus the right-hand side is equal to $a_{11} \det A_{11}$. If A is a lower triangular matrix, then for each $i = 2, \dots, n$, in the first row of the matrix A_{i1} there are only zero entries, hence $\det A_{i1} = 0$ according to the remark following Theorem 15. In this case the right-hand side is equal to $a_{11} \det A_{11}$ as well. Due to the induction assumption, we obtain

$$\det A = a_{11} \det A_{11} = a_{11} \cdot a_{22} \cdot \dots \cdot a_{nn}.$$

∎

Theorem 17 (determinant and invertibility)**.** Let $A \in M(n \times n)$. Then A is invertible if and only if $\det A \neq 0$.

Proof. If A is invertible, then we can transform it to the identity matrix. We know $\det I = 1 \neq 0$ and thus according to Theorem 15(iv) $\det A \neq 0$.

If A is not invertible, then $h(A) < n$ (Theorem 11), and thus we can transform A to an upper triangular matrix A', which has at least one zero entry on the diagonal. Then $\det A' = 0$ and hence $\det A = 0$. ∎

Theorem 18 (determinant of matrix product)**.** For matrices $A, B \in M(n \times n)$, the following formula $\det AB = \det A \cdot \det B$ holds.

Proof. Let us put $C = AB$. We divide the proof into two cases.

1. Let us assume first that the matrix A is not invertible. Let us choose a transformation T which alter the matrix A to A' of a row echelon form. Since a transformation does not change a rank and $h(A) < n$, then the last row of the matrix A' is a zero vector. Let C' denote a matrix which becomes from C by the transformation T. According to Theorem 9 $C' = A'B$ holds, therefore the last row of the matrix C' is also a zero vector. From that follows $h(C') < n$. The matrix C has the same rank and hence it is not invertible. We obtain $\det A = \det C = 0$ and the equality is proved.

2. Now let us suppose that the matrix A is invertible. Then we can tranform it to the identity matrix. Let us take the transformation T which alter the identity matrix I to A. Since $IB = B$ and $AB = C$, from Theorem 9 it follows that the matrix C becomes from B by the transformation T. From Theorem 15 it follows that there exists such a number $\alpha \in \mathbb{R}$ that if D is $n \times n$ matrix and a matrix D' becomes from the matrix D by the transformation T, then $\det D' = \alpha \det D$. Thus, in a special case $\det A = \alpha \det I = \alpha$. Next, we have

$$\det C = \alpha \det B = \det A \cdot \det B,$$

and thus the proof is completed. ∎

Remark. An assertion similar to Theorem 15 holds also for column elementary operations. We can use this in proving the two following theorems - but we will omit their proofs.

Theorem 19 (determinant and transposition)**.** Let $A \in M(n \times n)$. Then $\det A = \det A^T$.

Theorem 20. Let $A = (a_{is})_{\substack{i=1..n \\ s=1..n}} \in M(n \times n)$ and $j \in \{1, \ldots, n\}$. Then

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij},$$

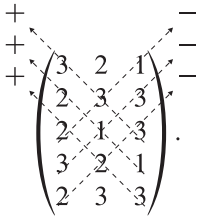
$$\det A = \sum_{s=1}^{n} (-1)^{s+j} a_{js} \det A_{js}.$$

The first (the second) formula is called an expansion of the determinant along the j-th column (row, respectively).

Example 21. Calculate the determinant of the matrix

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

Solution. We can calculate the determinant of 3×3 matrix by so-called **Sarrus' rule**, which can be easily proved from definition. We first add two more rows to the matrix A such that the fourth row of the new matrix is the first row of the matrix A and the fifth row of the new matrix is the second row of the matrix A. This way we get a 5×3 matrix



Now we calculate the determinant of the matrix A by adding the products of the entries on diagonals denoted by the sign $+$ together and from that number we subtract the products of the entries on diagonals denoted by the sign $-$, i.e.

$$\det A = 3 \cdot 3 \cdot 3 + 2 \cdot 1 \cdot 1 + 2 \cdot 2 \cdot 3 - 1 \cdot 3 \cdot 2 - 3 \cdot 1 \cdot 3 - 3 \cdot 2 \cdot 2 = 14.$$

♣

Example 22. Calculate the determinant of the matrix

$$A = \begin{pmatrix} 1 & -1 & 2 & 4 \\ 0 & 1 & -1 & 2 \\ 3 & -1 & 2 & 0 \\ -1 & 0 & 3 & 2 \end{pmatrix}.$$

Solution. We calculate the determinants of $n \times n$ matrix, where $n \geq 4$, by using the expansion along an arbitrary row (column). It is suitable to choose a row or column with the most zero entries.

Let us expanse the determinant along the first column:

$$\det A = 1 \cdot (-1)^{1+1} \cdot \begin{vmatrix} 1 & -1 & 2 \\ -1 & 2 & 0 \\ 0 & 3 & 2 \end{vmatrix} + 0 \cdot (-1)^{2+1} \cdot \begin{vmatrix} -1 & 2 & 4 \\ -1 & 2 & 0 \\ 0 & 3 & 2 \end{vmatrix} +$$
$$+ 3 \cdot (-1)^{3+1} \begin{vmatrix} -1 & 2 & 4 \\ 1 & -1 & 2 \\ 0 & 3 & 2 \end{vmatrix} + (-1) \cdot (-1)^{4+1} \cdot \begin{vmatrix} -1 & 2 & 4 \\ 1 & -1 & 2 \\ -1 & 2 & 0 \end{vmatrix}.$$

We can use the Sarrus' rule to calculate each of the 3×3 determinants – we get $\det A = 48$.

In order not to calculate as much determinants of 3×3 matrices, it is suitable to transform the matrix before the calculation by the row elementary operations of the third type (which does not change the determinant value) to a matrix which has

in one column only one non-zero entry. In case of our matrix A, we can proceed this way: We copy the first and the second row, then we subtract three times the first row and add the first row to the fourth row. That gives us the matrix

$$\begin{pmatrix} 1 & -1 & 2 & 4 \\ 0 & 1 & -1 & 2 \\ 0 & 2 & -4 & -12 \\ 0 & -1 & 5 & 6 \end{pmatrix}.$$

From the expansion along the first column and the Sarus' rule, we get

$$\det A = 1 \cdot (-1)^{1+1} \cdot \begin{vmatrix} 1 & -1 & 2 \\ 2 & -4 & -12 \\ -1 & 5 & 6 \end{vmatrix} = 48.$$

Another way of calculation is to transform the matrix $A \in M(n \times n)$ by the row elementary operations of the first and the third type to the upper triangular matrix A'. The determinant of the matrix A' is equal to the product of the entries on the main diagonal (see Theorem 16). The determinant of the matrix A is then equal to $(-1)^p \det A'$, where p denotes the number of row elementary operations of the first type which we used in transformation of the matrix A to the matrix A'. Since, while using row elementary operations of the first type, we need to consider the change of a sign of a determinant. This way is suitable for big n and we can take an advantage of using it in numerical calculations. ♣

2.4. Solving systems of linear equations

Let us consider a system of m linear equations in n variables $x_1, x_2, \ldots, x_n$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m, \end{aligned} \tag{S}$$

where $a_{ij} \in \mathbb{R}$, $b_i \in \mathbb{R}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

If we put

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

then we can write the given system (S) in a vector form $A\vec{x} = \vec{b}$. We call the matrix A the **matrix of the system** (S) and the vector $\vec{b}$ the **vector of the right-hand sides**. We call the matrix

$$(A|\vec{b}) = \left(\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{array}\right)$$

the **augmented matrix of the system** (S).

Gaussian elimination. We show a method how to solve the systems of linear equations effectively. Let $A \in M(m \times n)$ and $\vec{b} \in M(m \times 1)$. Here and subsequently we assume that the matrix A is non-zero. In the opposite case the problem is trivial. We can transform the matrix A by a transformation T to a matrix A' in a row echelon form (Theorem 6(i)). Let us apply T also to the $\vec{b}$ and let the result by denoted by $\vec{b}'$. Then for $\vec{y} \in M(n \times 1)$ follows $A\vec{y} = \vec{b}$ if and only if $A'\vec{y} = \vec{b}'$ (Theorem 9 and Theorem 6(ii)), in other words, the systems $A\vec{x} = \vec{b}$ and $A'\vec{x} = \vec{b}'$ have the same solution set.

Now we distinguish two cases according to the last non-zero row of the matrix $(A'|\vec{b}')$. Let the k-th row be the last non-zero row, $k \in \{1, \dots, m\}$.

1. If the k-th row of the matrix A' is a zero vector, then the k-th equation of the system $A'\vec{x} = \vec{b}'$ is of the form

$$0 \cdot x_1 + \dots + 0 \cdot x_n = b'_k \quad (\neq 0)$$

and thus the system has evidently no solution.

2. Now we suppose that the k-th row of the matrix A' is a non-zero vector. Let j_p, $p = 1, \dots, k$, be the smallest natural number such that $a'_{pj_p} \neq 0$. The entry a'_{pj_p} is thus the first non-zero entry in the p-th row. There exists such an entry, since the matrix A' is in a row echelon form and $p \leq k$. Let us put $I_1 = \{j_1, \dots, j_k\}$ and then $I_2 = \{1, \dots, n\} \setminus I_1$. Let us write the system $A'\vec{x} = \vec{b}'$ as

$$\sum_{s \in I_1} a'_{is} x_s + \sum_{s \in I_2} a'_{is} x_s = b'_i, \quad i = 1, \dots, k,$$

which can be rewritten in the form

$$\begin{aligned}
a'_{1j_1}x_{j_1} &= b'_1 - \sum_{s\in I_2} a'_{1s}x_s - a'_{1j_2}x_{j_2} - a'_{1j_3}x_{j_3} - \cdots - a'_{1j_k}x_{j_k},\\
a'_{2j_2}x_{j_2} &= b'_2 - \sum_{s\in I_2} a'_{2s}x_s - a'_{2j_3}x_{j_3} - \cdots - a'_{2j_k}x_{j_k},\\
&\vdots\\
a'_{k-1,j_{k-1}}x_{j_{k-1}} &= b'_{k-1} - \sum_{s\in I_2} a'_{k-1,s}x_s - a'_{k-1,j_k}x_{j_k},\\
a'_{kj_k}x_{j_k} &= b'_k - \sum_{s\in I_2} a'_{ks}x_s.
\end{aligned} \tag{5}$$

If we choose the numbers x_s, $s \in I_2$ arbitrarily, then we can calculate uniquely x_t, $t \in I_1$, from the system (5): We calculate from the last (i.e. the k-th) equation the variable x_{j_k}. We plug this number together with the numbers x_s, $s \in I_2$, into the $(k-1)$-st equation and solve the equation for $x_{j_{k-1}}$. Now we repeat the procedure in an obvious way, until we finally solve the first equation for x_{j_1}. The solution set of the system (5) (and thus also of the system (S)) consist of the vectors $\vec{y} = (y_1, \ldots, y_n)^T$, where the values y_s, $s \in I_2$, are chosen arbitrarily and the values y_s, $s \in I_1$, are determined by the system (5).

Let us show this procedure in the following example.

Example 23. Find all solutions of the system $A\vec{x} = \vec{b}$, where

$$A = \begin{pmatrix} 1 & 4 & 3 & 5 & 4\\ 2 & 5 & 3 & 7 & 5\\ 1 & 3 & 2 & 4 & 3\\ 0 & 1 & 1 & 2 & 2 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} 2\\ 1\\ 1\\ 0 \end{pmatrix}.$$

Solution. We transform the augmented matrix of the system $(A|\vec{b})$ by row elementary operations to the matrix in a row echelon form

$$\left(\begin{array}{ccccc|c} 1 & 2 & 1 & 2 & 1 & 1\\ 0 & 1 & 1 & 1 & 1 & 1\\ 0 & 0 & 0 & 1 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right). \tag{6}$$

It can be seen from the matrix form that the given system has a solution. The matrix (6) corresponds with the system

$$\begin{aligned}
x_1 + 2x_2 + x_3 + 2x_4 + x_5 &= 1,\\
x_2 + x_3 + x_4 + x_5 &= 1,\\
x_4 + x_5 &= -1.
\end{aligned} \tag{7}$$

If we use the notation from the previous part, then it follows $I_1 = \{1, 2, 4\}$ and $I_2 = \{3, 5\}$. We could thus rewrite the system in the form

$$\begin{aligned} x_1 &= 1 - 2x_2 - x_3 - 2x_4 - x_5, \\ x_2 &= 1 - x_3 - x_4 - x_5, \\ x_4 &= -1 - x_5. \end{aligned} \tag{8}$$

Now we determine the set of all solutions of the system this way: We can choose the values x_3 and x_5 arbitrarily. Thus, let us put $x_3 = s \in \mathbb{R}$ and $x_5 = t \in \mathbb{R}$. Then we can calculate consecutively from the system (8):

$$\begin{aligned} x_4 &= -1 - t, \\ x_2 &= 1 - s - (-1 - t) - t = 2 - s, \\ x_1 &= 1 - 2(2 - s) - s - 2(-1 - t) - t = -1 + s + t. \end{aligned}$$

The set of all solutions of the system is then

$$\big\{(-1 + s + t, 2 - s, s, -1 - t, t)^T;\ s, t \in \mathbb{R}\big\}.$$

♣

From the method of solution of the system described above and the fact that $h(A) = h(A')$ and $h(A|\vec{b}) = h(A'|\vec{b'})$, we get the following theorem.

Theorem 24 (Rouché-Fontené)**.** The system (S) has a solution if and only if the matrix of the system has the same rank as the augmented matrix of this system.

Let us now consider a special case, where the matrix of the system (S) is square and invertible above that. Let us examine what we get from the above method in this case. Applying the Gaussian elimination, we obtain $k = n$, $I_1 = \{1, \dots, n\}$ and $I_2 = \emptyset$. The system (5) is thus of the form

$$\begin{aligned} a'_{11}x_1 &= b'_1 - a'_{12}x_2 - \cdots - a'_{1n}x_n, \\ &\vdots \\ a'_{n-1,n-1}x_{n-1} &= b'_{n-1} - a'_{n-1,n}x_n, \\ a'_{nn}x_n &= b'_n. \end{aligned}$$

It can be seen that we can calculate the values $x_1, \dots, x_n$ uniquely and the system (S) thus has exactly one solution. The relation between the solubility of the system of linear equations and invertibility of its matrix is specified in the following theorem.

Theorem 25. Let $A \in M(n \times n)$. Then the conditions (i)–(iii) are equivalent:

(i) the matrix A is invertible,

(ii) the system (S) has exactly one solution for each $\vec{b} \in M(n \times 1)$,

(iii) the system (S) has at least one solution for each $\vec{b} \in M(n \times 1)$.

Proof. (i) $\Rightarrow$ (ii) If A is an invertible matrix, then the equation $A\vec{x} = \vec{b}$ has only one solution $\vec{x} = A^{-1}\vec{b}$.

(ii) $\Rightarrow$ (iii) Obvious.

(iii) $\Rightarrow$ (i) We prove that non (i) $\Rightarrow$ non (iii) holds. If A is not invertible, then $h(A) < n$ and thus there exists a transformation T_1, which alter A to a matrix S in a row echelon form with a zero vector in the last row. If we put $\vec{c} = (1, \ldots, 1)^T \in M(n \times 1)$, we get $h(S|\vec{c}) > h(S)$. Let T_2 be a transformation from Theorem 6(ii) and let $\vec{b} \in M(n \times 1)$ be such that $\vec{c} \overset{T_2}{\rightsquigarrow} \vec{b}$. Since $S \overset{T_2}{\rightsquigarrow} A$ and $(S|\vec{c}) \overset{T_2}{\rightsquigarrow} (A|\vec{b})$, then according to Theorem 6(iii) $h(A|\vec{b}) = h(S|\vec{c}) > h(S) = h(A)$ holds. According to Theorem 24, the system (S) has no solution. ∎

We show one more possibility how to describe the solution of a system, whose matrix is invertible. But this result has a rather theoretical (see the chapters **??** and **??**) than practical meaning , since its usage is usually more demanding for calculation then Gaussian elimination.

Theorem 26 (Cramer's rule)**.** Let $A \in M(n \times n)$ be an invertible matrix, $\vec{b} \in M(n \times 1)$, $\vec{x} \in M(n \times 1)$ and $A\vec{x} = \vec{b}$. Then

$$x_j = \frac{\begin{vmatrix} a_{11} & \ldots & a_{1,j-1} & b_1 & a_{1,j+1} & \ldots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \ldots & a_{n,j-1} & b_n & a_{n,j+1} & \ldots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & \ldots & a_{1,j-1} & a_{1j} & a_{1,j+1} & \ldots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \ldots & a_{n,j-1} & a_{nj} & a_{n,j+1} & \ldots & a_{nn} \end{vmatrix}}$$

for $j = 1, \ldots, n$.

Proof. Let $\vec{x}$ be a solution of the matrix $A\vec{x} = \vec{b}$. Then:

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + x_j \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}.$$

We can rewrite the previous equality to

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + 1 \cdot \begin{pmatrix} x_j a_{1j} - b_1 \\ \vdots \\ x_j a_{nj} - b_n \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

From that we obtain that the matrix

$$B = \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ x_j a_{1j} - b_1 & \dots & x_j a_{nj} - b_n \\ \dots & \dots & \dots \\ a_{1n} & \dots & a_{nn} \end{pmatrix}$$

has a rank smaller than n, so $\det B = 0$. From Theorems 14 and 15(i) we obtain

$$\det B = \begin{vmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ x_j a_{1j} & \dots & x_j a_{nj} \\ \dots & \dots & \dots \\ a_{1n} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ -b_1 & \dots & -b_n \\ \dots & \dots & \dots \\ a_{1n} & \dots & a_{nn} \end{vmatrix} =$$

$$= x_j \begin{vmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ a_{1j} & \dots & a_{nj} \\ \dots & \dots & \dots \\ a_{1n} & \dots & a_{nn} \end{vmatrix} - \begin{vmatrix} a_{11} & \dots & a_{n1} \\ \dots & \dots & \dots \\ b_1 & \dots & b_n \\ \dots & \dots & \dots \\ a_{1n} & \dots & a_{nn} \end{vmatrix} = 0.$$

According to Theorem 19 it follows that

$$0 = \det B = x_j \begin{vmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{vmatrix} - \begin{vmatrix} a_{11} & \dots & b_1 & \dots & a_{1n} \\ a_{21} & \dots & b_2 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & b_n & \dots & a_{nn} \end{vmatrix}.$$

From the last equation, we can easily deduce the required equation. ∎

Example 27. Solve the system of the equations

$$\begin{aligned} x_1 + x_2 - x_3 - x_4 &= 0, \\ x_1 + 2x_2 - x_3 + x_4 &= 5, \\ 2x_1 - x_2 + x_3 + 2x_4 &= 1, \\ -x_1 + x_2 + x_3 - x_4 &= 4. \end{aligned}$$

Solution. We use the method of Gaussian elimination on the augmented matrix of the system.

$$\left(\begin{array}{cccc|c} 1 & 1 & -1 & -1 & 0 \\ 1 & 2 & -1 & 1 & 5 \\ 2 & -1 & 1 & 2 & 1 \\ -1 & 1 & 1 & -1 & 4 \end{array}\right).$$

1. Let us copy the first row, subtract the first row from the second row, subtract double of the first row from the third row and add the first row to the fourth row.

We obtain

$$\left(\begin{array}{cccc|c} 1 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 2 & 5 \\ 0 & -3 & 3 & 4 & 1 \\ 0 & 2 & 0 & -2 & 4 \end{array}\right).$$

2. Let us copy the first and the second row, add three times the second row to the third row, subtract double of the second row from the fourth row. The originated matrix is of the form

$$\left(\begin{array}{cccc|c} 1 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 2 & 5 \\ 0 & 0 & 3 & 10 & 16 \\ 0 & 0 & 0 & -6 & -6 \end{array}\right).$$

By this transformation we obtain the system

$$\begin{aligned} x_1 + x_2 - x_3 - x_4 &= 0, \\ x_2 + 2x_4 &= 5, \\ 3x_3 + 10x_4 &= 16, \\ -6x_4 &= -6, \end{aligned}$$

which has the same solution as the original system.

From the last equation we obtain $x_4 = 1$, by plugging it into the third equation we get $x_3 = 2$; similarly, from the second equation we get $x_2 = 3$ and from the first $x_1 = 0$. The vector $(0, 3, 2, 1)^T$ is thus the only solution of the system. ♣

Remark. If we have an objective to solve the systems with the same invertible matrix of the system A, but for several right-hand sides $\vec{b}$, it could be more advantageous to find a matrix A^{-1} and get the solution by multiplying the system $A\vec{x} = \vec{b}$ by the inversion matrix on the left: $\vec{x} = A^{-1} \cdot A\vec{x} = A^{-1}\vec{b}$.

In our example it is

$$A^{-1} = \begin{pmatrix} 1/2 & -1/9 & 1/3 & 1/18 \\ 0 & 1/3 & 0 & 1/3 \\ 0 & -1/9 & 1/3 & 5/9 \\ -1/2 & 1/3 & 0 & -1/6 \end{pmatrix}.$$

The solution for general right-hand side $\vec{b}$ is then of the form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = A^{-1} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} b_1/2 - b_2/9 + b_3/3 + b_4/18 \\ b_2/3 + b_4/3 \\ -b_2/9 + b_3/3 + 5b_4/9 \\ -b_1/2 + b_2/3 - b_4/6 \end{pmatrix}.$$

And especially for $(b_1, b_2, b_3, b_4)^T = (0, 5, 1, 4)^T$ we have the same solution

$$(x_1, x_2, x_3, x_4)^T = (0, 3, 2, 1)^T.$$

Example 28. Solve the system of the linear equations

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= -2, \\ 3x_1 - 8x_2 - 2x_3 &= 4, \\ x_1 + 4x_3 &= -2. \end{aligned}$$

Solution. Let us use the Gaussian elimination:

$$\left(\begin{array}{ccc|c} -1 & 2 & 1 & -2 \\ 3 & -8 & -2 & 4 \\ 1 & 0 & 4 & -2 \end{array}\right), \qquad \left(\begin{array}{ccc|c} -1 & 2 & 1 & -2 \\ 0 & -2 & 1 & -2 \\ 0 & 2 & 5 & -4 \end{array}\right), \qquad \left(\begin{array}{ccc|c} -1 & 2 & 1 & -2 \\ 0 & -2 & 1 & -2 \\ 0 & 0 & 6 & -6 \end{array}\right).$$

From that we easily get the result: $x_3 = \frac{1}{6}(-6) = -1$, $x_2 = -\frac{1}{2}(-2+1) = \frac{1}{2}$, $x_1 = -(-2+1-2\cdot\frac{1}{2}) = 2$. ♣

Example 29. Solve the system of the equations

$$\begin{aligned} 2(a-1)x + (3a+1)y + az &= 2a, \\ (1-a)x - 2y - z &= 2, \\ ax + 2ay + az &= a+1 \end{aligned}$$

in dependence on a real parameter a.

Solution. As we know, the solubility of the system is related to the invertibility of its matrix A. If the matrix is invertible ($\det A \neq 0$), then the system has exactly one solution. Let us then calculate

$$\det A = \begin{vmatrix} 2(a-1) & 3a+1 & a \\ 1-a & -2 & -1 \\ a & 2a & a \end{vmatrix} = a(a+1)(a-2).$$

1. If $a \neq 0, a \neq -1$ and $a \neq 2$, then the system has exactly one solution and it is

$$x = \frac{1+3a}{a(2-a)}, \quad y = \frac{a+1}{a}, \quad z = \frac{a^2-4a-3}{a(2-a)}.$$

We could easily get the result by using the Cramer's rule (Theorem 26).

2. If $a = 0$, then the third equation of the system is $0 = 1$ and thus the system has no solution in this case.

3. For $a = -1$ the augmented matrix of the system is of the form

$$\left(\begin{array}{ccc|c} -4 & -2 & -1 & -2 \\ 2 & -2 & -1 & 2 \\ -1 & -2 & -1 & 0 \end{array}\right)$$

and we transform it to the matrix

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & -6 & -3 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

For $a = -1$ are thus the rank of the matrix and also of the augmented matrix equal to 2 and the original system is equivalent to the system

$$\begin{aligned} x + 2y + z &= 0, \\ -6y - 3z &= 2, \end{aligned}$$

which has infinitely many solutions: $x = 2/3$, $y = t$, $z = -2/3 - 2t$, where t is an arbitrary real number.

4. For $a = 2$ the augmented matrix of the system is

$$\left(\begin{array}{ccc|c} 2 & 7 & 2 & 4 \\ -1 & -2 & -1 & 2 \\ 2 & 4 & 2 & 3 \end{array}\right).$$

We transform it by row elementary operations to the matrix of the form

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & -2 \\ 0 & 3 & 0 & 8 \\ 0 & 0 & 0 & 7 \end{array}\right),$$

from this it can be seen that the rank of the matrix of the system, which is equal to 2, differs from the rank of the augmented matrix of the system, which equals 3. The system thus has no solution for $a = 2$. ♣

2.5. Matrices and linear mapping

Now, we will deal with linear mappings. Let us start with definition describing the notion of a linear mapping.

Definition. We say, that the mapping $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is **linear**, if:

(i) $\forall \vec{u}, \vec{v} \in \mathbb{R}^n\colon f(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v})$,
(ii) $\forall \lambda \in \mathbb{R}\ \forall \vec{u} \in \mathbb{R}^n\colon f(\lambda \vec{u}) = \lambda f(\vec{u})$.

Remark. In this section we will consider the elements of the space $\mathbb{R}^n$ as column vectors, i.e. as $n \times 1$ matrices.

Let $i \in \{1, \ldots, n\}$. We call the following vector with n entries

$$\vec{e^i} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ \\ i\text{-th coordinate} \\ \\ \\ \\ \end{matrix}$$

the ***i*-th canonical basis vector** of the space $\mathbb{R}^n$. We call the set of all canonical basis vectors in $\mathbb{R}^n$ **canonical basis** of the space $\mathbb{R}^n$. The canonical basis has two very important properties, which follow easily from the definitions:

(i) $\forall \vec{x} \in \mathbb{R}^n \colon \vec{x} = x_1 \cdot \vec{e}^1 + \cdots + x_n \cdot \vec{e}^n$,
(ii) the vectors $\vec{e}^1, \ldots, \vec{e}^n$ are linearly independent.

Theorem 30 (representation of linear mappings)**.** The mapping $f \colon \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there exists a matrix $A \in M(m \times n)$ such that

$$\forall \vec{u} \in \mathbb{R}^n \colon f(\vec{u}) = A\vec{u}.$$

Proof. $\Rightarrow$ Let the i-th coordinate of the vector $\vec{v} \in \mathbb{R}^m$ be denoted by $(\vec{v})_i$. Let $f \colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping. Let us put $a_{ij} = \big(f(\vec{e}^j)\big)_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, then it is sufficient to show that $\big(f(\vec{u})\big)_i = (A\vec{u})_i$, $i = 1, \ldots, m$, for each $\vec{u} \in \mathbb{R}^n$. Let us calculate

$$\big(f(\vec{u})\big)_i = \left(f\left(\sum_{j=1}^{n} u_j \vec{e}^j\right)\right)_i = \sum_{j=1}^{n} u_j \big(f(\vec{e}^j)\big)_i = \sum_{j=1}^{n} u_j a_{ij} = (A\vec{u})_i.$$

$\Leftarrow$ This implication follows from the definition of a matrix multiplication and from the theorem about matrix multiplication properties (Theorem 3). ∎

Remark. The matrix A from the previous theorem is determined uniquely. If $f(\vec{u}) = A\vec{u}$ should hold for each vector $\vec{u} \in \mathbb{R}^n$, then it must follow in a special case $f(\vec{e}^j) = A\vec{e}^j$ for each $j \in \{1, \ldots, n\}$. At the same time $A\vec{e}_j$ is the j-th column of the matrix A. It shows that the matrix A is determined uniquely (it must have the vector $f(\vec{e}^j)$ in the j-th column) and explains why the matrix A was defined in a certain way.

We call the matrix A from the previous theorem the **representing matrix** of a mapping f or that the matrix **represents** a mapping f.

Example 31. From the previous theorem it follows that all linear mappings from $\mathbb{R}$ to $\mathbb{R}$ are of the form $x \mapsto ax$, where $a \in \mathbb{R}$. Thus, these are the well known linear functions.

Similarly, linear mappings from $\mathbb{R}^2$ to $\mathbb{R}$ are of the form $L \colon \vec{x} \mapsto a_1 x_1 + a_2 x_2$, where $a_1, a_2 \in \mathbb{R}$. Let us note, that a graph of the function L is a plane in $\mathbb{R}^3$ passing through the origin and $\nabla L(\vec{x}) = (a_1, a_2)$ holds for each $\vec{x} \in \mathbb{R}^2$. See the following figure.

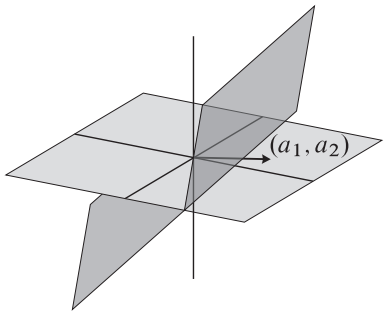


FIGURE 1.

Example 32. Examples of the linear mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$ are an anisotropic dilatation or a rotation through an angle φ. The representing matrix A is given by

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}, \quad \text{respectively.}$$

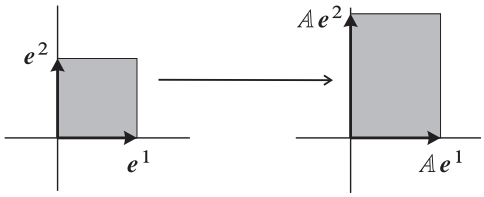


FIGURE 2. An anisotropic dilatation in $\mathbb{R}^2$

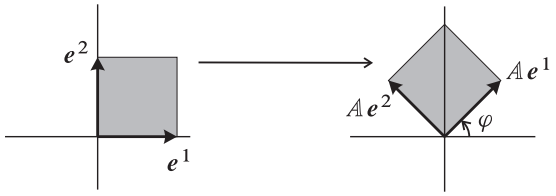


FIGURE 3. A rotation through an angle φ

Similarly, an example of a linear mapping from $\mathbb{R}^3$ to $\mathbb{R}^3$ is a dilatation representing by the matrix

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

or a rotation through an angle φ around the axis given by a vector $\vec{v}^1 \in \mathbb{R}^3$.

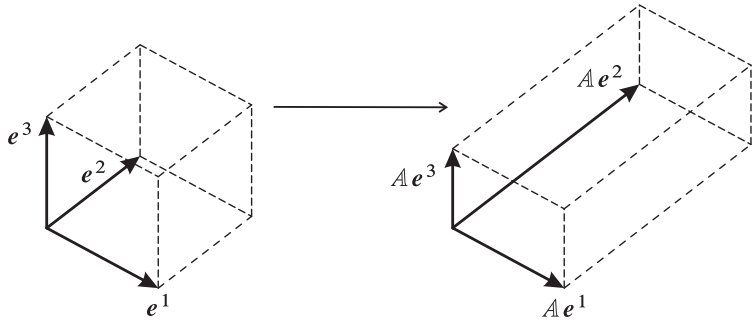


FIGURE 4. Anisotropic dialatation in $\mathbb{R}^3$

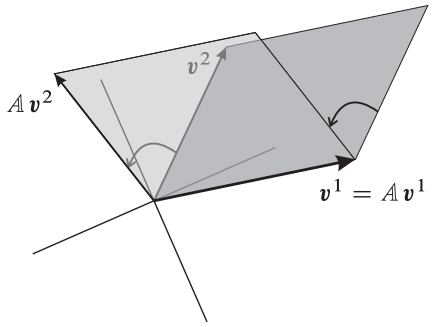


FIGURE 5. Rotation in $\mathbb{R}^3$

Theorem 33. Let a mapping $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be linear. Then the following conditions are equivalent:

(i) f is a bijection (that is f is an one-to-one and an onto mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$),
(ii) f is an ono-to-one mapping,
(iii) f is a surjective mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$.

Proof. (i) $\Rightarrow$ (ii) This implication is obvious.

(ii) $\Rightarrow$ (iii) We prove it by a contradiction. Let thus f be injective, but not surjective. Let A be a matrix representing the mapping f. The matrix A is not invertible (Theorem 25) and thus A^T is not invertible (Theorem 5). The rows of the matrix A^T are thus linearly dependent. Since the rows of the matrix A^T are

the columns of the matrix A (let them be denoted by $\vec{s}^1, \ldots, \vec{s}^n$), there exists their non-trivial linear combination which is equal to a zero vector:

$$x_1\vec{s}^1 + \cdots + x_n\vec{s}^n = \vec{o}.$$

We could rewrite the last equality: $A\vec{x} = \vec{o}$, where $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Then we have $A\vec{x} = \vec{o}$, $\vec{x} \neq \vec{o}$ and also $A\vec{o} = \vec{o}$. This contradicts the assumption that f is an injective mapping.

(iii) $\Rightarrow$ (i) This implication follows from Theorem 25 about solving linear equations systems. ∎

Theorem 34 (composition of linear mappings)**.** Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping representing by a matrix $A \in M(m \times n)$ and $g\colon \mathbb{R}^m \to \mathbb{R}^k$ be a linear mapping representing by a matrix $B \in M(k \times m)$. Then the composite mapping $g \circ f\colon \mathbb{R}^n \to \mathbb{R}^k$ is linear and represented by the matrix BA.

Proof. For $\vec{v} \in \mathbb{R}^n$ it follows that

$$(g \circ f)(\vec{v}) = g\big(f(\vec{v})\big) = g(A\vec{v}) = B(A\vec{v}) = (BA)\vec{v}.$$

We used the asociativity of a matrix multiplication here. ∎

Remarks. 1. The previous theorem shows us the relation between a composition of linear mappings and a matrix multiplication. The representing matrix of linear mappings composition is equal to the product of their representing matrices in corresponding order.

2. Let a linear mapping $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a bijection. It can be easily justified that an inversion mapping f^{-1} is also linear. Let an identity mapping on $\mathbb{R}^n$ be denoted by Id. If f is represented by a matrix A and f^{-1} is represented by a matrix B, then due to the relation $f \circ f^{-1} = f^{-1} \circ f = \mathrm{Id}$ $AB = BA = I$ holds (according to Theorem 34). Thus $B = A^{-1}$, in other words the mapping f^{-1} is represented by the matrix A^{-1}.

3. Let $AB = I$ holds for matrices $A, B \in M(n \times n)$. If we take a linear mapping $f\colon \mathbb{R}^n \to \mathbb{R}^n$ represented by a matrix A and a linear mapping $g\colon \mathbb{R}^n \to \mathbb{R}^n$ represented by a matrix B, then $f \circ g = \mathrm{Id}$ holds. Hence, g is an injection and f is a surjection. According to Theorem 33 f and g are bijections. From that it follows that g is an inversion mapping to f. According to the previous points $B = A^{-1}$ holds, especially we have $BA = I$. We proved that from relation $AB = I$ follows necessarily $BA = I$ provided $A, B \in M(n \times n)$.

Example 35. Let a mapping $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ be defined by

$$f(u_1, u_2, u_3) = (u_1 - u_2, u_1 - 2u_2, u_1 - 3u_2)^T.$$

Show that f is a linear mapping and determine its representing matrix.

Solution. Let $\vec{u} = (u_1, u_2, u_3)^T \in \mathbb{R}^3$ and $\vec{v} = (v_1, v_2, v_3)^T \in \mathbb{R}^3$. Then

$$\begin{aligned} f(\vec{u} + \vec{v}) = \\ &= (u_1 + v_1 - u_2 - v_2, u_1 + v_1 - 2u_2 - 2v_2, u_1 + v_1 - 3u_2 - 3v_2)^T = \\ &= (u_1 - u_2, u_1 - 2u_2, u_1 - 3u_2)^T + (v_1 - v_2, v_1 - 2v_2, v_1 - 3v_2)^T = \\ &= f(\vec{u}) + f(\vec{v}). \end{aligned}$$

Let λ be an arbitrary real number. Then

$$\begin{aligned} f(\lambda\vec{u}) &= (\lambda u_1 - \lambda u_2, \lambda u_1 - 2\lambda u_2, \lambda u_1 - 3\lambda u_2)^T = \\ &= \lambda(u_1 - u_2, u_1 - 2u_2, u_1 - 3u_2)^T = \lambda f(\vec{u}). \end{aligned}$$

We checked that the mapping f has both properties from the definition of a linear mapping and thus it is linear. From the function formula for f it can be easily seen that

$$f(\vec{u}) = f(u_1, u_2, u_3) = \begin{pmatrix} 1 & -1 & 0 \\ 1 & -2 & 0 \\ 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}.$$

But we can also realize that the columns of the representing matrix are the vectors $f(\vec{e}^1)$, $f(\vec{e}^2)$ and $f(\vec{e}^3)$, see the proof of Theorem 30. ♣

By the end of this chapter let us show one more usage of determinants.

Example 36. Calculate the area of a triangle ABC, whose vertices has the coordinates $A = [1, 1, 0], B = [3, 0, 2], C = [0, -1, 1]$ in $\mathbb{R}^3$.

Solution. It is known from geometry that the area of a triangle ABC can be calculated according to the formula $p = \frac{1}{2}bc\sin\alpha$, where b is the lenght of a line segment AC, c is the lenght of a line segment AB and an angle α is an interior angle of the triagle at a vertex A. For the calculation of the values $b, c, \sin\alpha$ we use the vectors $\vec{u} = B - A$ and $\vec{v} = C - A$. It is in fact $b = \|\vec{v}\|$, $c = \|\vec{u}\|$ and it can be proved that $\cos\alpha = \frac{\vec{u}\vec{v}}{\|\vec{u}\|\|\vec{v}\|}$. (Let the symbol $\vec{u}\vec{v}$ denote the **scalar product** of vectors $\vec{u} = (u_1, u_2, u_3)^T$ and $\vec{v} = (v_1, v_2, v_3)^T$, which is defined by: $\vec{u}\vec{v} = u_1v_1 + u_2v_2 + u_3v_3$. We use the symbol $\|\vec{u}\| = \sqrt{u_1^2 + u_2^2 + u_3^2}$ to denote the **length of the vector** $\vec{u}$.) The angle α from the interval $(0, \pi)$ is uniquely determined by the expression for cosinus of this angle. We can calculate easily $\sin\alpha = \sqrt{1 - \cos^2\alpha}$.

In our case we have $\vec{u} = (2, -1, 2)^T$, $\vec{v} = (-1, -2, 1)^T$, $\|\vec{u}\| = 3$, $\|\vec{v}\| = \sqrt{6}$, $\cos\alpha = \sqrt{6}/9$ and $\sin\alpha = 5\sqrt{3}/9$. The area p of the triangle ABC is thus equal to $p = 5\sqrt{2}/2$.

We can use another interesting formula to calculate the area of a triangle in $\mathbb{R}^3$. Let us define the vector product $\vec{u} \times \vec{v}$ of two vectors $\vec{u} = (u_1, u_2, u_3)^T$ and $\vec{v} =$

$(v_1, v_2, v_3)^T$ by

$$\vec{u} \times \vec{v} = \left(\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}, - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \right)^T .$$

By a direct computation we can immediatelly find out that $\vec{u}\vec{u} \times \vec{v} = \vec{v}\vec{u} \times \vec{v} = 0$, in other words the vector product of two vectors is perpendicular to each of them.

We can easily calculate from the definition of the vector product

$$\|\vec{u} \times \vec{v}\| = \sqrt{\|\vec{u}\|^2 \|\vec{v}\|^2 - \vec{u}\vec{v}^2} = \|\vec{u}\| \|\vec{v}\| \sqrt{1 - \cos^2 \alpha} = \|\vec{u}\| \|\vec{v}\| \sin \alpha,$$

from that we obtain another way how to express the area of the triangle ABC, namely $p = \frac{1}{2} \|\vec{u} \times \vec{v}\|$. ♣

Remark. Let us derive a formula for volume of a parallelepiped $ABCDA'B'C'D'$. We know that the volume is given by the formula $V = P \cdot h$, where P is the area of the base $ABCD$ and h is the length of the two basis $ABCD$ and $A'B'C'D'$. Let us put $\vec{u} = B - A$, $\vec{v} = D - A$, $\vec{w} = A' - A$. According to what we derived previously, we can write $P = \|\vec{u} \times \vec{v}\|$.

Now we need to calculate h. We know that the vector $\vec{u} \times \vec{v}$ is perpendicular to a plane of the lower base $ABCD$. Let θ denote the angle between vectors $\vec{w}$ and $\vec{u} \times \vec{v}$, then we get $|\cos \theta| = h / \|\vec{w}\|$ and thus $h = |\cos \theta| \|\vec{w}\|$.

Thus we obtain $V = \|\vec{u} \times \vec{v}\| \|\vec{w}\| |\cos \theta|$. But we have at the same time:

$$|\cos \theta| = \frac{|\vec{w}\vec{u} \times \vec{v}|}{\|\vec{w}\| \|\vec{u} \times \vec{v}\|}.$$

By plugging the expression into the formula for V, we finally get

$$V = |\vec{w}\vec{u} \times \vec{v}| = \left| \det \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \right| .$$

Thus we can summarize this: the volume of a parallelepiped which is determined by three linearly independent vectors $\vec{u}$, $\vec{v}$, $\vec{w}$ in $\mathbb{R}^3$ is equal to the absolute value of the determinant of the matrix

$$\begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} .$$

2.6. Cvičení

1. Calculate the product AB, where

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ -2 & 3 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

2. Calculate the products AB and BA, where

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

3. Calculate the product AB, where

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 4 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 7 \\ 5 & 2 & 1 \end{pmatrix}.$$

4. Determine a matrix X so that the equality $AX = B$ holds, provided

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 3 & 5 \\ 5 & 9 \end{pmatrix}.$$

5. Calculate $A^n = \underbrace{AA\cdots A}_{n\text{-times}}$ for each $n \in \mathbb{N}$, if

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

6. Solve the matrix equations system

$$3X + 2Y = 12A, \qquad 4X + 3Y = 17A$$

with unknown matrices X and Y, where

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 2 & 0 & -1 & 1 \end{pmatrix}.$$

7. Determine the rank $h(A)$ of the matrix

$$A = \begin{pmatrix} -1 & 2 & -3 & 5 & 1 \\ 1 & -1 & 5 & -2 & -1 \\ -2 & 5 & -2 & 10 & 1 \\ 0 & 1 & 4 & 0 & -1 \\ -1 & 3 & 1 & 5 & 0 \end{pmatrix}.$$

8. Determine the rank $h(A)$ of the matrix

$$A = \begin{pmatrix} 4 & 0 & 1 & 1 \\ 9 & 1 & 2 & 2 \\ 2 & -2 & a & 1 \\ a & 0 & a+1 & 3 \end{pmatrix}$$

in dependance on a parameter $a \in \mathbb{R}$.

9. Find the inverse of the matrix

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 0 & -2 \\ 2 & 1 & 1 \end{pmatrix}.$$

10. Find the inverse of the matrix

$$A = \begin{pmatrix} 1 & 0 & -1 & 2 \\ 1 & 1 & -1 & 1 \\ -2 & 0 & 3 & -6 \\ -4 & -1 & 6 & -10 \end{pmatrix}.$$

11. Find the inverse of a product AB of the matrices

$$A = \begin{pmatrix} 1 & -1 & 2 & 0 & 3 \\ -2 & 1 & 0 & -1 & 1 \\ 3 & 0 & -1 & 2 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 2 & -4 \\ 0 & 6 & -2 \\ 2 & -1 & 2 \\ 1 & -4 & 9 \\ -1 & 1 & 0 \end{pmatrix}.$$

In the following four exercises find the determinant of a given matrix.

12. $\begin{pmatrix} 3 & -8 \\ 2 & 5 \end{pmatrix}$

13. $\begin{pmatrix} 2 & -1 & 2 \\ 3 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

14. $\begin{pmatrix} 1 & 0 & 2 & 3 \\ 2 & 1 & 3 & 2 \\ 4 & 0 & 3 & -1 \\ 3 & 2 & 1 & -2 \end{pmatrix}$

15. $\begin{pmatrix} 2 & 2 & 1 & 0 & -1 \\ 2 & 1 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 2 \\ -1 & 0 & 1 & 2 & 2 \end{pmatrix}$

Solve the following equations in $\mathbb{R}$.

16. $\begin{vmatrix} x & 1 & 0 \\ 2 & -1 & 1 \\ 1 & -2x & -2 \end{vmatrix} = 9$

17. $\begin{vmatrix} 5-x & 6 & -3 \\ 6 & 9-x & 0 \\ -3 & 0 & 9-x \end{vmatrix} = 0$

Solve the systems of equations / the systems of equations with parameter $a \in \mathbb{R}$.

18.

$$\begin{aligned} 4x + 3y + 2z &= 1 \\ x + 3y + 5z &= 1 \\ 3x + 6y + 9z &= 2 \end{aligned}$$

19.

$$\begin{aligned} 5x - y + 2z &= 1 \\ 3x + 5y - z &= 2 \\ 2x - 6y + 3z &= 4 \end{aligned}$$

20.

$$\begin{aligned} x_1 - x_2 - 3x_4 &= -1 \\ 7x_1 - 2x_2 + 2x_3 - 10x_4 &= -2 \\ 7x_1 - x_2 + x_3 - 9x_4 &= -4 \\ 2x_1 - 2x_3 - 4x_4 &= -6 \\ 6x_1 - x_2 + 2x_3 - 7x_4 &= -1 \end{aligned}$$

21.

$$\begin{aligned} ax + y + z &= 1 \\ x + ay + z &= a \\ x + y + az &= a^2 \end{aligned}$$

22.

$$\begin{aligned} x - 5y - 7z &= 0, \\ -2x + y + az &= -3, \\ -x + ay + 3z &= -1 \end{aligned}$$

23.

$$\begin{aligned} x + 2y + 3z + 4t &= 1, \\ 2x - 2y + 3z - 3t &= -5, \\ x + y + z + t &= 5, \\ 4x + 3y - 5z + 2t &= 3 \end{aligned}$$

24. Solve the system of equations

$$\begin{aligned} x_1 + x_2 - x_3 &= 5, \\ x_1 - 4x_2 + 2x_3 &= -1, \\ x_1 - x_2 + x_3 &= 1 \end{aligned}$$

by using an inversion matrix (if it exists).

25. Find all solutions of the system $A\vec{x} = \vec{b}$, where

$$A = \begin{pmatrix} 1 & 3 & 2 & 4 \\ 1 & 1 & 0 & 3 \\ -1 & -3 & -2 & -2 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} -1 \\ -4 \\ -1 \end{pmatrix}.$$

26. Find all solutions of the system $A\vec{x} = \vec{b}$, where

$$A = \begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 2 & 3 & 1 & 5 & 3 \\ 1 & 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} 0 \\ -2 \\ -2 \\ 3 \end{pmatrix}.$$

27. Find all solutions of the system $A\vec{x} = \vec{b}$, where

$$A = \begin{pmatrix} 1 & 3 & 2 & 4 & 3 & -2 \\ 1 & 1 & 0 & 3 & 2 & -2 \\ -1 & -3 & -2 & -2 & -1 & -3 \\ 0 & 1 & 1 & 0 & 0 & -1 \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} -4 \\ -7 \\ -3 \\ 1 \end{pmatrix}.$$

28. Let mappings f, g and h from $\mathbb{R}^3$ to $\mathbb{R}^4$ be given by:

(i) $f(u_1, u_2, u_3) = (u_1 + u_2 + u_3, u_1 + u_2 - u_3, u_1 - u_2 + u_3, -u_1 + u_2 + u_3)^T$,
(ii) $g(u_1, u_2, u_3) = (u_1, 2u_2, -u_3 + 5u_1, 0)^T$,
(iii) $h(u_1, u_2, u_3) = (1, u_1, u_2, u_3)^T$.

Examine in each case if the mapping is linear. If so, determine its representing matrix.

29. Let $f\colon \mathbb{R}^4 \to \mathbb{R}^3$ be a linear mapping determined by a matrix A and $g\colon \mathbb{R}^3 \to \mathbb{R}^4$ be a linear mapping determined by a matrix B, where

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 1 & -3 & -1 & -1 \\ 0 & 0 & 3 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 2 & -1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Determine a matrix of a mapping $g \circ f\colon \mathbb{R}^4 \to \mathbb{R}^4$ and a matrix of a mapping $f \circ g\colon \mathbb{R}^3 \to \mathbb{R}^3$. Write formulas of these mappings.

Results of exercises

1. $AB = \begin{pmatrix} x + 2y + 3z \\ y - z \\ -2x + 3y \end{pmatrix}$ **2.** $AB = \begin{pmatrix} 3 & 1 \\ 3 & -3 \end{pmatrix}$, $BA = \begin{pmatrix} -2 & 2 \\ 4 & 2 \end{pmatrix}$ **3.** $AB = \begin{pmatrix} 20 & 11 & 12 \\ 18 & 8 & 30 \end{pmatrix}$ **4.** $X = \begin{pmatrix} -1 & -1 \\ 2 & 3 \end{pmatrix}$ **5.** It can be proved by mathematical induction that $A^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$ holds. **6.** $X = 2A = \begin{pmatrix} 2 & 2 & 2 & 0 \\ 4 & 0 & -2 & 2 \end{pmatrix}$, $Y = 3A = \begin{pmatrix} 3 & 3 & 3 & 0 \\ 6 & 0 & -3 & 3 \end{pmatrix}$ **7.** $h(A) = 3$ **8.** $h(A) = 4$, if $a \neq 1$ and $a \neq 12$; $h(A) = 3$ for $a = 1, a = 12$ **9.** Inverse of the matrix does not exist, since $h(A) = 2$.

10.

$$A^{-1} = \begin{pmatrix} 3 & 0 & 1 & 0 \\ -2 & 2 & -2 & 1 \\ 0 & 2 & -3 & 2 \\ -1 & 1 & -2 & 1 \end{pmatrix}$$

11.

$$(AB)^{-1} = \begin{pmatrix} 22 & 8 & -5 \\ 5 & 2 & -1 \\ -3 & -1 & 1 \end{pmatrix}$$

12. 31 **13.** 11 **14.** -24 **15.** 8 **16.** The equation has two solutions: $x_1 = 1$, $x_2 = -2$. **17.** The equation has three solutions: $x_1 = 9$, $x_2 = 0$, $x_3 = 14$. **18.** The system has infinitely many solutions: $x = t$, $y = 1/3 - 2t$, $z = t$, $t \in \mathbb{R}$. **19.** The solution does not exists.

20. Given system has infinitely many solutions of the form $x_1 = -6/7 + 8t/7$, $x_2 = 1/7 - 13t/7$, $x_3 = 15/7 - 6t/7$, $x_4 = t$, where $t \in \mathbb{R}$.

21. For $a \neq 1, a \neq -2$ the system has one solution:

$$x = -\frac{a+1}{a+2}, \quad y = \frac{1}{a+2}, \quad z = \frac{(a+1)^2}{a+2};$$

for $a = 1$ the system has infinitely many solutions of the form:

$$x = 1 - s - t, \quad y = s, \quad z = t,$$

where $s, t \in \mathbb{R}$; for $a = -2$ the system has no solution.

22. For $a \neq 2$ and $a \neq 17$ the system has one solution:

$$x = \frac{-26}{a-17}, \quad y = \frac{-1}{a-17}, \quad z = \frac{-3}{a-17};$$

for $a = 2$ the system has infinitely many solutions:

$$x = 5/3 + t/3, \quad y = 1/3 - 4t/3, \quad z = t,$$

where $t \in \mathbb{R}$; for $a = 17$ the system has no solution.

23. $x = -3$, $y = 13$, $z = 2$, $t = -7$

24. The system has one solution:

$$\begin{pmatrix} 1/2 & 0 & 1/2 \\ -1/4 & -1/2 & 3/4 \\ -3/4 & -1/2 & 5/4 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ -2 \end{pmatrix}.$$

25. $x_1 = -3 + t$, $x_2 = 2 - t$, $x_3 = t$, $x_4 = -1$; $t \in \mathbb{R}$ **26.** $x_1 = -3 + t_1 + t_2$, $x_2 = 3 - t_1$, $x_3 = t_1$, $x_4 = -1 - t_2$, $x_5 = t_2$; $t_1, t_2 \in \mathbb{R}$ **27.** $x_1 = -2 + t_1 - t_2$, $x_2 = t_2$, $x_3 = 2 - t_2$, $x_4 = -1 - t_1$, $x_5 = t_1$, $x_6 = 1$; $t_1, t_2 \in \mathbb{R}$

28. (i) The mapping f is linear and

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

(ii) The mapping g is linear,

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 5 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

(iii) The mapping h is not linear.

29. The matrix of the mapping $g \circ f$ is

$$BA = \begin{pmatrix} 0 & 3 & 6 & 3 \\ 3 & -6 & -3 & -2 \\ 1 & -3 & 2 & 0 \\ 1 & 0 & 2 & 1 \end{pmatrix},$$

$$(g \circ f)(u_1, u_2, u_3, u_4) =$$
$$= (3u_2 + 6u_3 + 3u_4, 3u_1 - 6u_2 - 3u_3 - 2u_4, u_1 - 3u_2 + 2u_3, u_1 + 2u_3 + u_4).$$

The matrix of the mapping $f \circ g$ is

$$AB = \begin{pmatrix} 2 & 1 & 3 \\ -3 & -8 & 3 \\ 1 & 3 & 3 \end{pmatrix},$$

$$(f \circ g)(u_1, u_2, u_3) = (2u_1 + u_2 + 3u_3, -3u_1 - 8u_2 + 3u_3, u_1 + 3u_2 + 3u_3).$$

CHAPTER 3

Integral

3.1. Primitive function

In this chapter we will be concerned with somewhat opposite operation to differentiation. In other words, let a function f be given. Then we will look for a function F whose derivative is equal to the original function f. As we will see, this problem is usually more complicated than finding a derivative of a given function.

Definition. Let a function $f\colon I \to \mathbb{R}$ be given, where I is an non-empty open interval. We say that the function $F\colon I \to \mathbb{R}$ is a **primitive function of f over I** provided that for each $x \in I$ there exists $F'(x)$ for which $F'(x) = f(x)$ holds.

Remarks. 1. The process of finding a primitive function is sometimes called **integration**, the function f an **integrand** and the primitive function an **indefinite integral**. We are always looking for a primitive function of a given function on a non-empty open interval.

2. If F is a primitive function of the function f over I, then according to Theorem **??** the function F is continuous on I.

3. A primitive function of a given function f is not determined uniquely. If a function F is a primitive function of f over an interval I, then also a function $x \mapsto F(x) + c$, $x \in I$, where $c \in \mathbb{R}$ is a constant, is a primitive function of f over I. However, the following theorem ensures that a primitive function is determined uniquely "up to a constant".

Theorem 1. Let F and G be primitive functions of a function f over an open interval I. Then there exists $c \in \mathbb{R}$ such that $F(x) = G(x) + c$ for each $x \in I$.

Proof. Let us put $H(x) = F(x) - G(x)$, $x \in I$. Then for a derivative of H follows $H'(x) = f(x) - f(x) = 0$ for each $x \in I$. According to the Theorem about monotonicity and the sign of the derivative, the function H is constant on interval I, which would complete the proof. ∎

Remark. Let the function f has a primitive function F on an open interval I. According to a previous remark and the theorem it could be seen that we obtain a

primitive function of the function f over I by adding a suitable constant function to one primitive function F.

Let the set of all primitive functions of a function f be denoted by a symbol

$$\int f(x)\,\mathrm{d}x.$$

However, a problem arises now how to describe this set precisely and simply. We will use a notation

$$\int f(x)\,\mathrm{d}x \overset{c}{=} F(x), \quad x \in I,$$

which means that for an arbitrary primitive function G of the function f there exists a constant $c \in \mathbb{R}$ such that $G = F + c$ on the interval I.

Theorem 2. Let f have a primitive function F over an open interval I, a function g have a primitive function G over I and $\alpha, \beta \in \mathbb{R}$. Then the function $\alpha F + \beta G$ is a primitive function of $\alpha f + \beta g$ over I.

Proof. The assertion follows from the equation $(\alpha F + \beta G)' = \alpha f + \beta g$ which holds on the interval I. ∎

The following theorem is very important, however, it will be proved in the latter section.

Theorem 3 (the existence of a primitive function)**.** Let f be a continuous function on an open interval I. Then f has a primitive function over I.

Remark. It is not difficult to realize, that there is no primitive function of the function signum on the whole $\mathbb{R}$. Conversely, we can also find discontinuous functions which have primitive functions.

Primitive functions of some important functions. We can check the validity of the following formulas by differentiating:

- $\displaystyle\int x^n\,\mathrm{d}x \overset{c}{=} \frac{x^{n+1}}{n+1},\ x \in \mathbb{R}$ for $n \in \mathbb{Z}$, $n \geq 0$; $x \in (-\infty, 0)$ or $x \in (0, +\infty)$ for $n \in \mathbb{Z}$, $n < -1$,
- $\displaystyle\int x^\alpha\,\mathrm{d}x \overset{c}{=} \frac{x^{\alpha+1}}{\alpha+1},\ x \in (0, +\infty)$ for $\alpha \in \mathbb{R} \setminus \{-1\}$,
- $\displaystyle\int \frac{1}{x}\,\mathrm{d}x \overset{c}{=} \log|x|,\ x \in (-\infty, 0)$ or $x \in (0, +\infty)$,
- $\displaystyle\int e^x\,\mathrm{d}x \overset{c}{=} e^x,\ x \in \mathbb{R}$,
- $\displaystyle\int \sin x\,\mathrm{d}x \overset{c}{=} -\cos x,\ x \in \mathbb{R}$,

- $\int \cos x \,\mathrm{d}x \overset{c}{=} \sin x,\ x \in \mathbb{R},$
- $\int \frac{1}{\cos^2 x}\,\mathrm{d}x \overset{c}{=} \operatorname{tg} x,\ x \in (-\pi/2 + k\pi, \pi/2 + k\pi),\ k \in \mathbb{Z},$
- $\int \frac{-1}{\sin^2 x}\,\mathrm{d}x \overset{c}{=} \operatorname{cotg} x,\ x \in (k\pi, \pi + k\pi),\ k \in \mathbb{Z},$
- $\int \frac{1}{1+x^2}\,\mathrm{d}x \overset{c}{=} \operatorname{arctg} x,\ x \in \mathbb{R},$
- $\int \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x \overset{c}{=} \arcsin x,\ x \in (-1,1),$
- $\int -\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x \overset{c}{=} \arccos x,\ x \in (-1,1).$

Basic methods of calculation of primitive functions are described in Theorem 4 and in Theorem 8.

Theorem 4 (integration by substitution)**.**

(i) Let F be a primitive function of f over (a,b). Let φ be a function defined on (α,β) with values in the interval (a,b) which is differentiable at each point of the interval (α,β). Then

$$\int f\big(\varphi(x)\big)\varphi'(x)\,\mathrm{d}x \overset{c}{=} F\big(\varphi(x)\big) \text{ over } (\alpha,\beta).$$

(ii) Let the function φ be differentiable at each point of the interval (α,β) and the derivative is either positive at all points or negative at all points and $\varphi\big((\alpha,\beta)\big) = (a,b)$. Let the function f be defined on interval (a,b) and

$$\int f\big(\varphi(t)\big)\varphi'(t)\,\mathrm{d}t \overset{c}{=} G(t) \text{ over } (\alpha,\beta).$$

Then

$$\int f(x)\,\mathrm{d}x \overset{c}{=} G\big(\varphi^{-1}(x)\big) \text{ over } (a,b).$$

Proof. (i) The assertion follows from the theorem about the derivative of the composite function (Theorem **??**) which says in this case that for each $x \in (\alpha,\beta)$ the derivative is $\big(F(\varphi(x))\big)' = f\big(\varphi(x)\big)\varphi'(x)$.

(ii) According to the assumption, φ is either increasing on (α,β) or decreasing on (α,β). Thus, there exists φ^{-1}. For each $x \in (a,b)$ then follows:

$$\big(G(\varphi^{-1}(x))\big)' = f\big(\varphi(\varphi^{-1}(x))\big)\varphi'(\varphi^{-1}(x))\frac{1}{\varphi'(\varphi^{-1}(x))} = f(x).$$

We used the theorem about the derivative of a composite function and the theorem about the derivative of an inverse function. ∎

Example 5. Determine a primitive function of the function $g(x) = \dfrac{x}{\sqrt{2+5x^2}}$.

Solution. The given function is continuous on the whole $\mathbb{R}$, thus it has a primitive function on the whole $\mathbb{R}$. For calculation of $\int g(x)\,\mathrm{d}x$ we use the substitution "$t = 2+5x^2$", i.e. the function $\varphi\colon \mathbb{R} \to (0,+\infty)$, $\varphi(x) = 2+5x^2$, since we notice that $\varphi'(x) = 10x$ and thus

$$\int \frac{x}{\sqrt{2+5x^2}}\,\mathrm{d}x = \frac{1}{10}\int \frac{\varphi'(x)}{\sqrt{\varphi(x)}}\,\mathrm{d}x.$$

According to Theorem 4(i) we need to calculate

$$\frac{1}{10}\int \frac{1}{\sqrt{t}}\,\mathrm{d}t \overset{c}{=} \frac{1}{5}\sqrt{t}, \quad t \in (0,+\infty).$$

Thus, each function

$$x \mapsto \frac{1}{5}\sqrt{2+5x^2} + c,$$

where $c \in \mathbb{R}$ is an arbitrary constant, is thus a primitive function of the function g over $\mathbb{R}$. ♣

Example 6. Determine a primitive function of the function $g(x) = \dfrac{1}{\sqrt{8+6x-9x^2}}$.

Solution. The functione g is continuous on its domain $(-2/3, 4/3)$, and thus has a primitive function over the interval. We first manipulate the function g in the following way:

$$g(x) = \frac{1}{\sqrt{9-(3x-1)^2}} = \frac{1}{3}\frac{1}{\sqrt{1-\left(\frac{3x-1}{3}\right)^2}}.$$

Let us then calculate

$$\frac{1}{3}\int \frac{1}{\sqrt{1-(x-1/3)^2}}\,\mathrm{d}x.$$

This integral is similar to the integral

$$\int \frac{1}{\sqrt{1-t^2}}\,\mathrm{d}t \overset{c}{=} \arcsin t.$$

We use Theorem 4(i). If we put φ=$x \mapsto x - 1/3$ (its derivative equals to 1), then we get

$$\frac{1}{3}\int \frac{1}{\sqrt{1-(x-1/3)^2}}\,\mathrm{d}x = \frac{1}{3}\int \frac{\varphi'(x)}{\sqrt{1-\varphi^2(x)}}\,\mathrm{d}x.$$

According to Theorem 4(i) it is sufficient to calculate

$$\frac{1}{3}\int \frac{1}{\sqrt{1-t^2}}\,\mathrm{d}t \overset{c}{=} \frac{1}{3}\arcsin t, \quad t \in (-1,1),$$

hence

$$\frac{1}{3}\int \frac{1}{\sqrt{1-(x-1/3)^2}}\,\mathrm{d}t \overset{c}{=} \frac{1}{3}\arcsin(x-1/3), \quad x\in(-2/3,4/3).$$

♣

Example 7. Determine a primitive function of the function $f(x)=\sqrt{1-x^2}$.

Solution. Let us search the primitive function over the interval $(-1,1)$, which is the maximal open interval contained in the domain of the function f. Here we choose $\varphi(t)=\sin t$, where $t\in(-\pi/2,\pi/2)$. The function φ has a positive derivative on the interval $(-\pi/2,\pi/2)$ and it maps the interval $(-\pi/2,\pi/2)$ to $(-1,1)$. Next,

$$\int f\big(\varphi(t)\big)\varphi'(t)\,\mathrm{d}t = \int |\cos t|\cos t\,\mathrm{d}t = \int \cos^2 t\,\mathrm{d}t = \\ = \int\left(\frac{1}{2}+\frac{1}{2}\cos 2t\right)\mathrm{d}t \overset{c}{=} \frac{1}{2}t+\frac{1}{4}\sin 2t$$

holds on the interval $(-\pi/2,\pi/2)$. We can easily check the last equation by differentiating, eventually by using the theorem about integration by substitution once more. Then according to Theorem 4(ii) we obtain

$$\int f(x)\,\mathrm{d}x \overset{c}{=} \frac{1}{2}\varphi^{-1}(x)+\frac{1}{4}\sin\big(2\varphi^{-1}(x)\big) = \frac{1}{2}\arcsin x+\frac{1}{4}\sin(2\arcsin x)$$

on the interval $(-1,1)$. ♣

Remark. If we want to use Theorem 4(i) to calculate a primitive function to the function g, it is necessary to find functions f and φ such that $g=(f\circ\varphi)\cdot\varphi'$ holds. Often, the procedure is to choose the function φ at first and then assign the function f to it. In Examples 5 and 6 we replaced the expression $\varphi'(x)\,\mathrm{d}x$ with the expression $\mathrm{d}t$ and the rest of the integrand was then of the form $f\circ\varphi$. Formally, we substitute $\varphi(x)=t$ and $\varphi'(x)\,\mathrm{d}x=\mathrm{d}t$. The last relation, although has no mathematical meaning, is helpful in calculation.

In case, that we did not success in finding the function f in a previously mentioned way, but the derivative of the function φ is positive everywhere (or negative everywhere), we can proceed the following way. We replace the expression x with the expression $\varphi^{-1}(t)$ and the expression $\mathrm{d}x$ with the expression $(\varphi^{-1})'(t)\,\mathrm{d}t$, so we get the expression $\int g\big(\varphi^{-1}(t)\big)(\varphi^{-1})'(t)\,\mathrm{d}t$. The integrand is then the searched function f. In fact

$$f\big(\varphi(x)\big)\cdot\varphi'(x) = g\big(\varphi^{-1}(\varphi(x))\big)(\varphi^{-1})'(\varphi(x))\cdot\varphi'(x) = g(x)$$

holds, while the last equality follows from the theorem about the derivative of an inverse function (Theorem **??**).

Theorem 8 (integration by parts)**.** Let I be an open interval and functions f and g be continuous on I. Let F be a primitive function of f over I and G be a primitive function of g over I. Then

$$\int f(x)G(x)\,\mathrm{d}x = F(x)G(x) - \int F(x)g(x)\,\mathrm{d}x \text{ over } I. \tag{1}$$

Remark. The expression (1) is an equality of two sets of functions. The set on the right-hand side contains functions of the form $FG - Z$, where Z is an arbitrary primitive function of the function Fg over I.

Proof. It is sufficient to realize that the functions fG and Fg are continuous and thus have primitive functions (Theorem 3) and that $(FG)' = fG + Fg$ holds on the interval I. ∎

Example 9. Determine a primitive function of the function $\varphi(x) = \sqrt{x}\log^2 x$.

Solution. The function φ is continuous on its domain $(0, +\infty)$, thus it has a primitive function over this interval. We use integration by parts for calculation of $\int \varphi(x)\,\mathrm{d}x$ (Theorem 8).

Let us put $f(x) = \sqrt{x}$, $G(x) = \log^2 x$ and calculate

$$\int f(x)\,\mathrm{d}x \overset{c}{=} \frac{2}{3}\sqrt{x^3} \qquad \text{and} \qquad g(x) = G'(x) = 2\log x \cdot \frac{1}{x}.$$

We have

$$\int \sqrt{x}\log^2 x\,\mathrm{d}x = \frac{2}{3}\sqrt{x^3}\log^2 x - \frac{4}{3}\int \sqrt{x}\log x.$$

We use again integration by parts for calculation of the last integral. Let now $f(x) = \sqrt{x}$ and $G(x) = \log x$, then

$$\int f(x)\,\mathrm{d}x \overset{c}{=} \frac{2}{3}\sqrt{x^3} \qquad \text{and} \qquad g(x) = G'(x) = \frac{1}{x}.$$

Finally we have on the interval $(0, +\infty)$

$$\begin{aligned}\int \sqrt{x}\log^2 x\,\mathrm{d}x &= \frac{2}{3}\sqrt{x^3}\log^2 x - \frac{4}{3}\left(\frac{2}{3}\sqrt{x^3}\log x - \frac{2}{3}\int \sqrt{x}\,\mathrm{d}x\right) \overset{c}{=} \\ &\overset{c}{=} \frac{2}{3}\sqrt{x^3}\log^2 x - \frac{8}{9}\sqrt{x^3}\log x + \frac{8}{9}\cdot\frac{2}{3}\sqrt{x^3} = \\ &= \frac{2}{3}\sqrt{x^3}\left(\log^2 x - \frac{4}{3}\log x + \frac{8}{9}\right).\end{aligned}$$

♣

Example 10. Let $n \in \mathbb{N}$. Determine a primitive function of $\dfrac{1}{(1+x^2)^n}$ over $\mathbb{R}$.

Solution. Let us put $I_n = \displaystyle\int \frac{1}{(1+x^2)^n}\,\mathrm{d}x$ and calculate according to Theorem 8:

$$\begin{aligned}
I_n &= \int \underbrace{1}_{f} \cdot \underbrace{\frac{1}{(1+x^2)^n}}_{G}\,\mathrm{d}x = \\
&= \underbrace{x}_{F} \cdot \underbrace{\frac{1}{(1+x^2)^n}}_{G} - \int \underbrace{x}_{F} \cdot \underbrace{(-n)\frac{2x}{(1+x^2)^{n+1}}}_{g}\,\mathrm{d}x = \\
&= \frac{x}{(1+x^2)^n} + 2n\int \frac{x^2}{(1+x^2)^{n+1}}\,\mathrm{d}x = \\
&= \frac{x}{(1+x^2)^n} + 2n\int \frac{1+x^2-1}{(1+x^2)^{n+1}}\,\mathrm{d}x = \\
&= \frac{x}{(1+x^2)^n} + 2nI_n - 2nI_{n+1}.
\end{aligned}$$

From that we calculate

$$I_{n+1} = \frac{x}{2n(1+x^2)^n} + \frac{2n-1}{2n}I_n, \quad x \in \mathbb{R}, n \in \mathbb{N}.$$

Since $I_1 \overset{c}{=} \operatorname{arctg} x$, this recurrence formula enables us to determine I_n for each $n \in \mathbb{N}$. For example

$$\begin{aligned}
I_2 &\overset{c}{=} \frac{x}{2(1+x^2)} + \frac{1}{2}\operatorname{arctg} x, \\
I_3 &\overset{c}{=} \frac{x}{4(1+x^2)^2} + \frac{3x}{8(1+x^2)} + \frac{3}{8}\operatorname{arctg} x.
\end{aligned}$$

♣

Theorem 3 says that a continuous function on an open interval has always a primitive function. However, we cannot always express this primitive function by elementary functions – more precisely by a finite number of addition, subtraction, multiplication, division and composition eof elementary functions. This property has for example the function e^{-x^2}, however, the proof is not easy. Now we show some types of functions, which have not this difficulty. Basic class of these functions are the rational functions. We show also some other types of functions, whose integration is possible to transform to integration of rational functions by a suitable substitution.

We first introduce some facts from the algebra. Let us note that if we have a polynomial

$$P(x) = a_n x^n + \cdots + a_1 x + a_0,$$

that we can plug into a variable x complex numbers as well and that the values then will be also complex numbers. Thus, we could take each polynomial as a mapping from $\mathbb{C}$ to $\mathbb{C}$ as well. In the rest of this section, we will consider also a polynomials with complex coefficients, i.e. $a_0, \ldots, a_n \in \mathbb{C}$. Degree of this polynomial is defined by an obvious way. In what follows, let the degree of a polynomial P be denoted by the symbol $\operatorname{st} P$.

Lemma 11 (about polynomials division)**.** Let P and Q be two polynomials (generally with complex coefficients) and the polynomial Q is not equal to zero. Then there exist uniquely determined polynomials R and Z satisfying:

- $\operatorname{st} Z < \operatorname{st} Q$,
- $P(x) = R(x)Q(x) + Z(x)$ for all $x \in \mathbb{C}$.

If P and Q have real coefficients, then also R and Z have real coefficients.

Proof. We prove the existence of the polynomials R and Z by applying mathematical induction on a degree of P. If $\operatorname{st} P = -1$, i.e. P is equal to zero, then put $R = Z = 0$. Now let us assume that the assertion holds for all polynomials P of degree less then k. Let us have a polynomial P of degree $k \geq 0$. If $\operatorname{st} P < \operatorname{st} Q$, then put $R = 0$ and $Z = P$. Otherwise, we write $m = \operatorname{st} Q$ and denote by a_k and b_m the coefficient of the term x^k of the polynomial P and the coefficient of the term x^m of the polynomial Q, respectively. If we set

$$\tilde{P}(x) = P(x) - \frac{a_k}{b_m} x^{k-m} Q(x),$$

then we obtain $\operatorname{st} \tilde{P} < k$ and hence from the induction assumption there exist polynomials $\tilde{R}$ and Z such that $\tilde{P} = \tilde{R}Q + Z$ and $\operatorname{st} Z < \operatorname{st} Q$ holds. Now it suffices to put $R(x) = \frac{a_k}{b_m} x^{k-m} + \tilde{R}(x)$.

If the polynomials P and Q have real coefficients, then from the previous procedure it can be seen that also polynomials R and Z have real coefficients.

It remains to prove the uniqueness. Let us suppose that

$$P = R_1 Q + Z_1 = R_2 Q + Z_2$$

for any polynomials R_1, R_2, Z_1, Z_2 and at the same time $\operatorname{st} Z_1 < \operatorname{st} Q$ and $\operatorname{st} Z_2 < \operatorname{st} Q$. Then $0 = (R_1 - R_2)Q + Z_1 - Z_2$ holds. The polynomial $R_1 - R_2$ is necessarily equal to zero. Otherwise, $\operatorname{st}\big((R_1 - R_2)Q\big) \geq \operatorname{st} Q > \operatorname{st}(Z_1 - Z_2)$ must hold, which contradicts with the equality $(R_1 - R_2)Q + Z_1 - Z_2 = 0$. From that it follows that $R_1 = R_2$ and $Z_1 = Z_2$. ∎

Corollary 12. If P is a polynomial and $\lambda \in \mathbb{C}$ its **root** (i.e. $P(\lambda) = 0$), then there exists a polynomial R satisfying $P(x) = (x - \lambda)R(x)$ for all $x \in \mathbb{C}$.

Proof. Let us put $Q(x) = x - \lambda$. Then according to Lemma 11 there exist polynomials R and Z such that $P = RQ + Z$, where $\operatorname{st} Z < \operatorname{st} Q = 1$. The polynomial Z is thus constant. We have $0 = P(\lambda) = R(\lambda)(\lambda - \lambda) + Z(\lambda)$ and then $Z(\lambda) = 0$. From that follows that Z is equal to zero. ∎

Theorem 13 (factoring to root terms)**.** Let $P(x) = a_n x^n + \cdots + a_1 x + a_0$ be a polynomial of degree $n \in \mathbb{N}$. Then there exist numbers $x_1, \ldots, x_n \in \mathbb{C}$ such that

$$P(x) = a_n(x - x_1) \cdots (x - x_n), \quad x \in \mathbb{C}. \tag{2}$$

Proof. We use mathematical induction. For $n = 1$ the assertion is obvious, since $P(x) = a_1\big(x - (-\frac{a_0}{a_1})\big)$ and it suffices to put $x_1 = -\frac{a_0}{a_1}$. Let then $n \in \mathbb{N}$, $n > 1$, and the assertion holds for all polynomials of degree less than or equal to $n - 1$. According to the fundamental theorem of algebra (Theorem **??**) there exists a root $x_n \in \mathbb{C}$ of the polynomial P. Due to Corollary 12 $P(x) = (x - x_n)R(x)$ holds for some polynomial R. Let us note that $\operatorname{st} R = n - 1$ and above that the coefficient of x^{n-1} of the polynomial R is equal to a_n. Thus, according to the induction assumption there exists a factorization $R(x) = a_n(x - x_1) \cdots (x - x_{n-1})$. From that we obtain a required factorization of the polynomial P. ∎

Remarks. 1. For each polynomial, the factorization (2) is unique up to the order of the terms. There are all roots of the polynomial P among the numbers $x_1, \ldots, x_n$. From that it follows that a polynomial of degree $n \in \mathbb{N}$ has at most n different roots.

2. The assertion of Corollary 12 can be strengthen even more in the following way. If P is a non-zero polynomial and $\lambda \in \mathbb{C}$, then there exists exactly one $k \in \mathbb{N} \cup \{0\}$ and a uniquely determined polynomial R satisfying $P(x) = (x - \lambda)^k R(x)$ for all $x \in \mathbb{C}$ and $R(\lambda) \neq 0$.

Since if we have $P(x) = (x - \lambda)^k R(x)$ for a polynomial R and $k \in \mathbb{N} \cup \{0\}$, then the polynomial R is necessarily non-zero and $k \leq \operatorname{st} P$. We could thus find the biggest $k \in \mathbb{N} \cup \{0\}$, for which there exists a polynomial R satisfying $P(x) = (x - \lambda)^k R(x)$. From Corollary 12 it follows that $R(\lambda) \neq 0$, otherwise we get a contradiction with maximality of k.

Let us prove the uniqueness. Let us assume, that $P(x) = (x - \lambda)^l \tilde{R}(x)$ holds for any $l \in \mathbb{N} \cup \{0\}$ and a polynomial $\tilde{R}$ satisfying $\tilde{R}(\lambda) \neq 0$. Let us note that from the choice of k follows $l \leq k$. Then we get $(x - \lambda)^{k-l} R(x) = \tilde{R}(x)$ for $x \neq \lambda$ and from the continuity it follows that this relation is satisfied also for $x = \lambda$. It must be $k = l$, otherwise by substituting $x = \lambda$ we obtain $\tilde{R}(\lambda) = 0$ and that is contradiction. Then also $R = \tilde{R}$ holds and this would complete the proof.

Definition. Let P be a non-zero polynomial, $\lambda \in \mathbb{C}$ and $k \in \mathbb{N}$. We say that a number λ is the **root of the multiplicity *k*** of a polynomial P if there exists a polynomial R satisfying $R(\lambda) \neq 0$ and $P(x) = (x - \lambda)^k R(x)$ for all $x \in \mathbb{C}$.

Remark. From the remark antecendent the definition it follows that the multiplicity of a root is uniquely determined and is equal to the number of occurences of the number λ in the n-tuple $x_1, x_2, \ldots, x_n$ from Theorem 13.

Polynomials with real coefficients have the following important property.

Theorem 14. Let P be a polynomial with real coefficients and $\lambda \in \mathbb{C}$ is a root of the polynomial P of multiplicity k. Then also a complex conjugate $\overline{\lambda}$ is a root of the polynom P of multiplicity k.

Proof. We first show that $P(\lambda) = 0$ if and only if $P(\overline{\lambda}) = 0$. Let us assume that the polynomial P is of the form $P(x) = a_n x^n + \cdots + a_1 x + a_0$, where $a_j \in \mathbb{R}$, $j = 0, \ldots, n$. Then

$$\begin{aligned} P(\overline{\lambda}) &= a_n(\overline{\lambda})^n + \cdots + a_1\overline{\lambda} + a_0 = a_n\overline{\lambda^n} + \cdots + a_1\overline{\lambda} + a_0 = \\ &= \overline{a_n\lambda^n} + \cdots + \overline{a_1\lambda} + \overline{a_0} = \overline{a_n\lambda^n + \cdots + a_1\lambda + a_0} = \overline{P(\lambda)}, \end{aligned}$$

From that the foregoing assertion follows.

We prove the theorem by applying mathematical induction on the degree of P. If $\operatorname{st} P = 1$, then λ is real and thus the proposition holds. Let us assume that the proposition holds for all polynomials of degree less than or equal to $n \in \mathbb{N}$. Let P be a polynomial with real coefficients of degree $n+1$ and $\lambda \in \mathbb{C}$ a root of P. If $\lambda = \overline{\lambda}$, then the proposition is obvious. Let us suppose that $\lambda \neq \overline{\lambda}$. According to the first part of the proof, $\overline{\lambda}$ is also a root of P. According to Corollary 12 there exists a polynomial Q satisfying

$$P(x) = (x - \lambda)(x - \overline{\lambda})Q(x),$$

where $\operatorname{st} Q < \operatorname{st} P$ holds. If we put

$$R(x) = (x - \lambda)(x - \overline{\lambda}) = x^2 - (\lambda + \overline{\lambda})x + \lambda\overline{\lambda} = x^2 - (2\operatorname{Re}\lambda)x + |\lambda|^2,$$

then we obtain that the polynomial R has (similarly to the polynomial P) real coefficients. According to Lemma 11 the polynomial Q thus has real coefficients. If λ is not a root of Q, then according to the first part of the proof neither $\overline{\lambda}$ is a root of Q and both numbers λ and $\overline{\lambda}$ are thus roots of P of multiplicity 1. If λ is a root Q of multiplicity l, then according to the induction assumption $\overline{\lambda}$ is also a root of Q of multiplicity l. Hence λ and $\overline{\lambda}$ are roots of P of multiplicity $l + 1$. ∎

Theorem 15. Let $P(x) = a_n x^n + \cdots + a_1 x + a_0$ be a polynomial of degree n with real coefficients. Then there exist real numbers $x_1, \ldots, x_k$, $\alpha_1, \ldots, \alpha_l$, $\beta_1, \ldots, \beta_l$ and natural numbers $p_1, \ldots, p_k$, $q_1, \ldots, q_l$ such that

- $P(x) = a_n(x - x_1)^{p_1} \cdots (x - x_k)^{p_k}(x^2 + \alpha_1 x + \beta_1)^{q_1} \cdots (x^2 + \alpha_l x + \beta_l)^{q_l}$,
- none of the two polynomials $x - x_1, \ldots, x - x_k, x^2 + \alpha_1 x + \beta_1, \ldots, x^2 + \alpha_l x + \beta_l$ has a root in common,
- the polynomials $x^2 + \alpha_1 x + \beta_1, \ldots, x^2 + \alpha_l x + \beta_l$ have no real root.

Proof. Let $x_1, \dots, x_k$ be all real roots (each of them different to each other) of the polynomial P with multiplicities $p_1, \dots, p_k$ and $z_1, \dots, z_l$ be roots of the polynomial P with positive imaginary part with multiplicities $q_1, \dots, q_l$. Then according to Theorem 14 the numbers $\overline{z_1}, \dots, \overline{z_l}$ are also roots of P with multiplicities $q_1, \dots, q_l$. Thus we can write

$$P(x) = a_n(x-x_1)^{p_1}\cdots(x-x_k)^{p_k}(x-z_1)^{q_1}(x-\overline{z_1})^{q_1}\cdots(x-z_l)^{q_l}(x-\overline{z_l})^{q_l}.$$

Next $(x-z_i)(x-\overline{z_i}) = x^2 + (-z_i-\overline{z_i})x + z_i\overline{z_i}$ holds. Both numbers $-z_i-\overline{z_i}$, $z_i\overline{z_i}$ are real, and therefore we can put $\alpha_i = -z_i-\overline{z_i}$ and $\beta_i = z_i\overline{z_i}$. It can be easily justified that the required properties are satisfied. ∎

Theorem 16 (partial fraction decomposition)**.** Let P, Q be polynomials with real coefficients such that $\operatorname{st} P < \operatorname{st} Q$ and let

$$Q(x) = a_n(x-x_1)^{p_1}\cdots(x-x_k)^{p_k}(x^2+\alpha_1 x+\beta_1)^{q_1}\cdots(x^2+\alpha_l x+\beta_l)^{q_l}$$

be a polynomial decomposition Q from Theorem 15. Then there exist uniquely determined real numbers $A_1^1,\dots, A_{p_1}^1,\dots, A_1^k,\dots,A_{p_k}^k$, B_1^1, $C_1^1,\dots, B_{q_1}^1$, $C_{q_1}^1,\dots,$ B_1^l, $C_1^l,\dots, B_{q_l}^l$, $C_{q_l}^l$ such that

$$\begin{aligned}\frac{P(x)}{Q(x)} = {} & \frac{A_1^1}{(x-x_1)} + \cdots + \frac{A_{p_1}^1}{(x-x_1)^{p_1}} + \cdots + \frac{A_1^k}{(x-x_k)} + \cdots + \frac{A_{p_k}^k}{(x-x_k)^{p_k}} + \\ & + \frac{B_1^1 x + C_1^1}{(x^2+\alpha_1 x+\beta_1)} + \cdots + \frac{B_{q_1}^1 x + C_{q_1}^1}{(x^2+\alpha_1 x+\beta_1)^{q_1}} + \cdots + \\ & + \frac{B_1^l x + C_1^l}{(x^2+\alpha_l x+\beta_l)} + \cdots + \frac{B_{q_l}^l x + C_{q_l}^l}{(x^2+\alpha_l x+\beta_l)^{q_l}},\ x \in \mathbb{R}\setminus\{x_1,\dots,x_k\}.\end{aligned}$$

The proof of this theorem is difficult more formally then by an idea and we will omit it.

Now we have prepared everything to search for a primitive function of a rational function.

Algorithm for calculating the primitive function of a rational function. Let P and Q be polynomials. If we have to integrate a rational function P/Q, then we proceed this way:

In case that the degree of P is greater than or equal to the degree of Q, we divide the polynomial P by the polynomial Q (Lemma 11) and obtain a decomposition

$$\frac{P(x)}{Q(x)} = R(x) + \frac{Z(x)}{Q(x)},$$

where R, Z are polynomials and the degree of Z is less than or equal to Q. It is easy to find a primitive function of R. If the polynomial Z is non-zero and $\operatorname{st} P < \operatorname{st} Q$, it remains to find a primitive function to the rational function Z/Q

and P/Q, respetively, where the degree of a numerator is less then the degree of a denominator. We decompose this function to partial fractions according to the previous theorem. Then we integrate each of the partial fractions.

Let us show now how to do it: We integrate a partial fraction corresponding to a real root a in the following way:

$$\int \frac{1}{(x-a)^n}\,\mathrm{d}x \overset{c}{=} \begin{cases} \frac{1}{1-n}\frac{1}{(x-a)^{n-1}} & \text{over } (-\infty,a) \text{ and over } (a,+\infty) \text{ for } n>1, \\ \log|x-a| & \text{over } (-\infty,a) \text{ and over } (a,+\infty) \text{ for } n=1. \end{cases}$$

A partial fraction of the form

$$\frac{Bx+C}{(x^2+\alpha x+\beta)^q},$$

where $B, C, \alpha, \beta \in \mathbb{R}$, $q \in \mathbb{N}$ and the polynomial $x^2+\alpha x+\beta$ has no real root is integrated in this way:

$$\int \frac{Bx+C}{(x^2+\alpha x+\beta)^q}\,\mathrm{d}x = \frac{B}{2}\underbrace{\int \frac{2x+\alpha}{(x^2+\alpha x+\beta)^q}\,\mathrm{d}x}_{I_1} + \left(C-\frac{B\alpha}{2}\right)\underbrace{\int \frac{1}{(x^2+\alpha x+\beta)^q}\,\mathrm{d}x}_{I_2}.$$

We could solve the integrals I_1 and I_2 as follows:

$$I_1 \overset{c}{=} \begin{cases} \frac{1}{(1-q)(x^2+\alpha x+\beta)^{q-1}} & \text{over } \mathbb{R} \text{ for } q>1, \\ \log(x^2+\alpha x+\beta) & \text{over } \mathbb{R} \text{ for } q=1; \end{cases}$$

$$I_2 = \int \frac{1}{\left((x+\alpha/2)^2+\beta-\alpha^2/4\right)^q}\,\mathrm{d}x = \\ = \frac{1}{(\beta-\alpha^2/4)^q}\int \frac{1}{\left(\left(\frac{x+\alpha/2}{\sqrt{\beta-\alpha^2/4}}\right)^2+1\right)^q}\,\mathrm{d}x.$$

In the last manipulation we used the inequality $\beta-\alpha^2/4>0$, which follows from the assumption that the polynomial $x^2+\alpha x+\beta$ has no real root. The discriminant of the equation $x^2+\alpha x+\beta=0$ is therefore negative. By using the substitution $t=\frac{x+\alpha/2}{\sqrt{\beta-\alpha^2/4}}$ we get an integrand of the form

$$\frac{1}{(1+t^2)^q}.$$

Integration of this function was shown in Example 10.

Example 17. Determine a primitive function of the function

$$f(x) = \frac{x}{(x^2+2x+2)^2(x^2+2x-3)}.$$

Solution. First we determine the domain of the function f. The expression $x^2 + 2x + 2$ is always positive, $x^2 + 2x - 3$ can be decomposed and $x^2 + 2x - 3 = (x-1)(x+3)$ holds. Hence it could be seen that $D_f = \mathbb{R}\setminus\{-3, 1\}$. The function f is continuous on the whole D_f. It thus has a primitive function over each of the intervals $(-\infty, -3)$, $(-3, 1)$ and $(1, +\infty)$.

Since the polynomial in the numerator is of smaller degree than the polynomial in the denominator, we can decompose the function f to partial fractions on D_f. The decomposition is of the form

$$\begin{aligned}\frac{x}{(x^2+2x+2)^2(x-1)(x+3)} &= \\ &= \frac{Ax+B}{x^2+2x+2} + \frac{Cx+D}{(x^2+2x+2)^2} + \frac{E}{x-1} + \frac{F}{x+3}.\end{aligned} \tag{3}$$

By multiplying this equation by the denominator of the left-hand side, we obtain the equation

$$\begin{aligned}x = {} & (Ax+B)(x^2+2x+2)(x-1)(x+3)+ \\ & + (Cx+D)(x-1)(x+3)+ \\ & + E(x^2+2x+2)^2(x+3) + F(x^2+2x+2)^2(x-1),\end{aligned} \tag{4}$$

which holds for each $x \in \mathbb{R} \setminus \{-3, 1\}$. However, polynomials are continuous on $\mathbb{R}$, and therefore the equation (4) holds for each $x \in \mathbb{R}$. Now we have two ways to proceed:

a) We compare the coefficients of the corresponding powers of x on the left-hand and right-hand side of the equation (4).

$$\begin{aligned}
x^5: &\qquad 0 = A + E + F,\\
x^4: &\qquad 0 = 4A + B + 7E + 3F,\\
x^3: &\qquad 0 = 3A + 4B + C + 20E + 4F,\\
x^2: &\qquad 0 = -2A + 3B + 2C + D + 32E,\\
x^1: &\qquad 1 = -6A - 2B - 3C + 2D + 28E - 4F,\\
x^0: &\qquad 0 = -6B - 3D + 12E - 4F.
\end{aligned}$$

Thus we get a linear system of six equations in six variables.

b) We substitute six different numbers for x in (4) and we obtain again a linear system of six equations in six variables. The most advantageous is to sustitute the

numbers, for which some summand equals to 0 (i.e. real roots of the denominator of the original fraction – in our case the number -3 and 1).

We usually combine those two method in a suitable way. By substituting 1 and -3 in (4) consecutively we obtain $E = 1/100$ and $F = 3/100$. We plug these values into the linear system obtained in a). From the first equation we get $A = -1/25$, from the second $B = 0$, from the last $D = 0$ and finally from the fourth $C = -1/5$. Hence we have determined the coefficients of the decomposition (3), which thus is of the form

$$f(x) = -\frac{1}{25}\cdot\frac{x}{x^2+2x+2} - \frac{1}{5}\cdot\frac{x}{(x^2+2x+2)^2} + \frac{1}{100}\cdot\frac{1}{x-1} + \frac{3}{100}\cdot\frac{1}{x+3}.$$

Now it remains to calculate primitive functions to individual partial fractions.

$$\begin{aligned}
\int \frac{x}{x^2+2x+2}\,\mathrm{d}x &= \frac{1}{2}\int \frac{2x+2}{x^2+2x+2}\,\mathrm{d}x - \int \frac{1}{x^2+2x+2}\,\mathrm{d}x = \\
&= \frac{1}{2}\log(x^2+2x+2) - \int \frac{1}{(x+1)^2+1}\,\mathrm{d}x \overset{c}{=} \\
&\overset{c}{=} \frac{1}{2}\log(x^2+2x+2) - \operatorname{arctg}(x+1),\ x\in\mathbb{R}, \\
\int \frac{x}{(x^2+2x+2)^2}\,\mathrm{d}x &= \frac{1}{2}\int \frac{2x+2}{(x^2+2x+2)^2}\,\mathrm{d}x - \int \frac{1}{(x^2+2x+2)^2}\,\mathrm{d}x = \\
&= -\frac{1}{2}\frac{1}{x^2+2x+2} - \int \frac{1}{\left((x+1)^2+1\right)^2}\,\mathrm{d}x \overset{c}{=} \\
&\overset{c}{=} -\frac{1}{2}\frac{1}{x^2+2x+2} - \frac{1}{2}\frac{x+1}{x^2+2x+2} - \frac{1}{2}\operatorname{arctg}(x+1),\ x\in\mathbb{R}, \\
\int \frac{1}{x-1}\,\mathrm{d}x &\overset{c}{=} \log|x-1|\,,\ x\in(-\infty,1) \text{ and } x\in(1,+\infty), \\
\int \frac{1}{x+3}\,\mathrm{d}x &\overset{c}{=} \log|x+3|\,,\ x\in(-\infty,-3) \text{ and } x\in(-3,+\infty).
\end{aligned}$$

On each of the intervals $(-\infty,-3)$, $(-3,1)$ and $(1,+\infty)$ the primitive function of the function f is thus an arbitrary function of the form

$$\begin{aligned}
&-\frac{1}{50}\log(x^2+2x+2) + \frac{7}{50}\operatorname{arctg}(x+1) + \frac{1}{10}\frac{x+2}{x^2+2x+2} + \\
&\qquad + \frac{1}{100}\log|x-1| + \frac{3}{100}\log|x+3| + c, \quad \text{where } c\in\mathbb{R}.
\end{aligned}$$

♣

Some useful substitutions. A polynomial in two variables is a function $[u, v] \mapsto \sum_{i,j=0}^{n} a_{ij} u^i v^j$, where $a_{ij} \in \mathbb{R}$, $n \in \mathbb{N} \cup \{0\}$. A rational function in two variables is a ratio of two polynomials in two variables.

Let R be a rational function in two variables.

1. For integration of the function $R(\sin x, \cos x)$ we can use the following substitution to change the integration of the function to the integration of a rational function.

(i) If $R(\sin x, -\cos x) = -R(\sin x, \cos x)$, the substitution $\sin x = t$ can be used.
(ii) If $R(-\sin x, \cos x) = -R(\sin x, \cos x)$, the substitution $\cos x = t$ can be used.
(iii) If $R(-\sin x, -\cos x) = R(\sin x, \cos x)$, the substitution $\operatorname{tg} x = t$ can be used.
(iv) The substitution $\operatorname{tg}(x/2) = t$ can be used always.

2. For integration of the function $R\left(x, \sqrt[q]{\frac{ax+b}{cx+d}}\right)$, where $q \in \mathbb{N}$ and the numbers $a, b, c, d \in \mathbb{R}$ satisfy $ad - bc \neq 0$, the substitution $t = \sqrt[q]{\frac{ax+b}{cx+d}}$ can be used to change the integration of the function to the integration of a rational function.

3. Integration of the function $R\left(x, \sqrt{ax^2 + bx + c}\right)$ can be also changed to the integration of a rational function. Here we need to distinguish 3 cases.

(i) The polynomial $ax^2 + bx + c$ has a real root x_1 of multiplicity two. Then, if the task have to make a sense, it must be $a > 0$, then it can be written $\sqrt{ax^2 + bx + c} = \sqrt{a}\,|x - x_1|$. The integrated function is thus rational on each of the intervals $(-\infty, x_1)$, $(x_1, +\infty)$.
(ii) The polynomial $ax^2 + bx + c$ has two real roots $x_1 < x_2$. Then it could be written $ax^2 + bx + c = a(x - x_1)(x - x_2)$. If $a > 0$, then we have

$$\sqrt{a(x - x_1)(x - x_2)} = \sqrt{a}\,|x - x_1| \sqrt{\frac{x - x_2}{x - x_1}}.$$

This equation shows that the function $R\left(x, \sqrt{ax^2 + bx + c}\right)$ could be on the intervals $(-\infty, x_1)$, $(x_2, +\infty)$ written of the form from the case 2. We can proceed similarly if $a < 0$.
(iii) The polynomial $ax^2 + bx + c$ has no real roots. If the task have to make a sense, it must be $a > 0$ or $c > 0$. Then we could use **Euler substitutions**

$$\sqrt{ax^2 + bx + c} = t + \sqrt{a}x \quad \text{or} \quad \sqrt{ax^2 + bx + c} = xt + \sqrt{c}.$$

These substitutions can be used also in the case (ii) provided that $a > 0$ and $c > 0$, respectively.

Example 18. Determine a primitive function of $f(x) = \dfrac{1}{1+3\cos^2 x}$.

Solution. The function f is continuous on the whole $\mathbb{R}$ and it thus has a primitive function over $\mathbb{R}$. If we put

$$R(u,v) = \frac{1}{1+3v^2},$$

then $f(x) = R(\sin x, \cos x)$. The equality $R(-\sin x, -\cos x) = R(\sin x, \cos x)$ holds and thus the substitution $t = \operatorname{tg} x$ can be used for $x \in \left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right)$, $k \in \mathbb{Z}$. To use this substitution, let us calculate

$$\cos^2 x = \frac{1}{1+\operatorname{tg}^2 x} = \frac{1}{1+t^2}.$$

Next, from the equality $x = \operatorname{arctg} t$ we obtain $\mathrm{d}x = \frac{1}{1+t^2}\,\mathrm{d}t$ according to the remark on page 111. We integrate the given integral by substitution

$$\int \frac{1}{1+3\frac{1}{1+t^2}} \cdot \frac{1}{1+t^2}\,\mathrm{d}t = \int \frac{1}{4+t^2}\,\mathrm{d}t \overset{c}{=} \frac{1}{2}\operatorname{arctg}\frac{t}{2}, \quad t \in \mathbb{R}.$$

According to the theorem about integration by substitution thus follows

$$\int \frac{1}{1+3\cos^2 x}\,\mathrm{d}x \overset{c}{=} \frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right), \quad x \in \left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right),\ k \in \mathbb{Z}.$$

If we put $F(x) = \frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right)$, then the function F is a primitive function of f over each of the intervals $\left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right)$, $k \in \mathbb{Z}$. However, we are searching for a primitive function over whole $\mathbb{R}$. Each primitive function G of f over $\mathbb{R}$ is equal to $F + c_k$ over the interval $\left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right)$, where $k \in \mathbb{Z}$ and $c_k \in \mathbb{R}$ is a suitable constant. Since G is continuous and the equalitites

$$\lim_{x \to \frac{\pi}{2}+k\pi-} G(x) = \frac{\pi}{4} + c_k \qquad \text{and} \qquad \lim_{x \to \frac{\pi}{2}+k\pi+} G(x) = -\frac{\pi}{4} + c_{k+1}$$

hold, it must be $c_{k+1} = c_k + \frac{\pi}{2}$ pro $k \in \mathbb{Z}$. Hence $c_k = c_0 + k\frac{\pi}{2}$, $k \in \mathbb{Z}$ and each primitive function of f is thus of the form

$$G(x) = \begin{cases} \frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right) + c_0 + k\frac{\pi}{2} & \text{for } x \in (-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi), \\ \frac{\pi}{4} + c_0 + k\frac{\pi}{2} & \text{for } x = \frac{\pi}{2} + k\pi. \end{cases}$$

The function G was made in a following way. On each interval $\left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right)$ we added a suitable constant to the function F such that the resulting function would be continuous, see the figures. This procedure is called “sticking”.

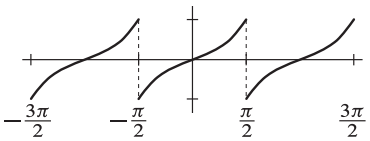


FIGURE 1.

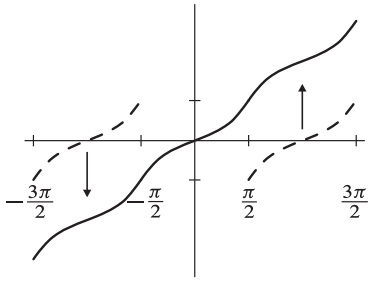


FIGURE 2.

♣

Example 19. Determine a primitive function of the function $f(x) = \dfrac{\sin x \cos x}{\sin^4 x + \cos^4 x}$.

Solution. The function f is continuous on $\mathbb{R}$, it thus has a primitive function over $\mathbb{R}$. Let us put

$$R(u, v) = \frac{uv}{u^4 + v^4}.$$

Then $f(x) = R(\sin x, \cos x)$ and it could be seen that:

(i) $R(\sin x, -\cos x) = -R(\sin x, \cos x)$. The substitution $t = \sin x$ can be used.

(ii) $R(-\sin x, \cos x) = -R(\sin x, \cos x)$. The substitution $t = \cos x$ can be used.

(iii) $R(-\sin x, -\cos x) = R(\sin x, \cos x)$. The substitution $t = \operatorname{tg} x$ can be used.

Certainly, we could also use the substitution $t = \operatorname{tg}(x/2)$, which is an universal substitution for transformation the integration of a rational function of sines and cosines to the integration of a rational function.

Let us try first the substitution $t = \operatorname{tg}(x/2)$ for $x \in (-\pi, \pi)$. To use this substitution, we calculate first

$$\cos x = \cos^2 \frac{x}{2} - \sin^2 \frac{x}{2} = \frac{\cos^2 \frac{x}{2} - \sin^2 \frac{x}{2}}{\cos^2 \frac{x}{2} + \sin^2 \frac{x}{2}} = \frac{1 - \operatorname{tg}^2 \frac{x}{2}}{1 + \operatorname{tg}^2 \frac{x}{2}} = \frac{1 - t^2}{1 + t^2},$$

$$\sin x = 2 \sin \frac{x}{2} \cos \frac{x}{2} = \frac{2 \sin \frac{x}{2} \cos \frac{x}{2}}{\cos^2 \frac{x}{2} + \sin^2 \frac{x}{2}} = \frac{2t}{1 + t^2},$$

$$\mathrm{d}x = \frac{2}{1 + t^2}\,\mathrm{d}t.$$

For determining $\mathrm{d}x$ we used the equality $x = 2 \operatorname{arctg} t$ and a remark on page 111.

By substitution, we manipulate the given integral to the integral

$$\int \frac{\frac{2t}{1+t^2}\cdot\frac{1-t^2}{1+t^2}}{\left(\frac{2t}{1+t^2}\right)^4+\left(\frac{1-t^2}{1+t^2}\right)^4}\cdot\frac{2}{1+t^2}\,\mathrm{d}t = 4\int\frac{t(1-t^2)(1+t^2)}{16t^4+(1-t^2)^4}\,\mathrm{d}t.$$

It can be seen, that we achieved our goal. However, the resulting function is complicated and above that we would have to overcome difficulties thatour substitution is valid only for $x\in(-\pi,\pi)$, eventually on an interval which is shifted by $2k\pi$, $k\in\mathbb{Z}$. Thus, let us try another substitutions.

1. The substitution $t=\sin x$. In our case we could manipulate the function f to the form

$$f(x)=\frac{\sin x}{\sin^4 x+(1-\sin^2 x)^2}\cdot\cos x.$$

If we realize that $\mathrm{d}t=\cos x\,\mathrm{d}x$ holds for the given substitution, we obtain

$$\int\frac{t}{2t^4-2t^2+1}\,\mathrm{d}t.$$

By using a substitution $u=t^2$ we then simplify the integrand and get

$$\int\frac{1}{4u^2-4u+2}\,\mathrm{d}u=\int\frac{1}{(2u-1)^2+1}\,\mathrm{d}u\overset{c}{=}\frac{1}{2}\operatorname{arctg}(2u-1),\ u\in\mathbb{R}.$$

The result is formed by the functions

$$\frac{1}{2}\operatorname{arctg}(2\sin^2 x-1)+c,\ x\in\mathbb{R},\quad c\in\mathbb{R}.$$

2. The substitution $t=\cos x$. We could manipulate the function f to the form

$$f(x)=\frac{\cos x}{(1-\cos^2 x)^2+\cos^4 x}\cdot\sin x.$$

If we realize that $\mathrm{d}t=-\sin x\,\mathrm{d}x$, we obtain

$$-\int\frac{t}{(1-t^2)^2+t^4}\,\mathrm{d}t\overset{c}{=}-\frac{1}{2}\operatorname{arctg}(2t^2-1),\ t\in\mathbb{R}.$$

(The calculation is analogous to the previous calculation.) The primitive function of f over $\mathbb{R}$ is every function of the form

$$-\frac{1}{2}\operatorname{arctg}(2\cos^2 x-1)+c,$$

where $c\in\mathbb{R}$ is an arbitrary constant.

3. Let us try one more substitution which can be used in our case – $t=\operatorname{tg}x$. We divide the numerator and the denominator in the formula of $f(x)$ by the expression $\cos^2 x$ and we get

$$f(x)=\frac{\operatorname{tg}x}{\sin^2 x\operatorname{tg}^2 x+\cos^2 x}=\frac{\operatorname{tg}x}{\operatorname{tg}^4 x+1}\cdot\frac{1}{\cos^2 x}.$$

Now we use the equality $\mathrm{d}t = \frac{1}{\cos^2 x}\,\mathrm{d}x$. Then we need to evaluate

$$\int \frac{t}{t^4+1}\,\mathrm{d}t \overset{c}{=} \frac{1}{2}\operatorname{arctg} t^2,\ t \in \mathbb{R}.$$

Thus we get the primitive function $\frac{1}{2}\operatorname{arctg}(\operatorname{tg}^2 x)$, but only over the intervals $(-\frac{\pi}{2}+k\pi, \frac{\pi}{2}+k\pi)$, $k \in \mathbb{Z}$. However, we know that the function f has a primitive function over the whole $\mathbb{R}$ (since it is continuous on $\mathbb{R}$). We could find this primitive function by the way described in the previous example.

Let us summarize: The substitution $t = \operatorname{tg} x$ was the easiest for computation. However, we did not get a primitive function over the whole D_f. We could solve this be "sticking" at the points which have to be omitted. By substituting $t = \operatorname{tg}(x/2)$ there is a similar situation – however, mostly leading to more complicated rational functions than in case of others substitutions. It is thus better – if the form of an integrand enables that – to avoid using it and to use some of the other three substitutions.

It could be seen from the foregoing that a form of the result could substantially depends on the substitution used, however, they are always functions which differ by a constant. ♣

Example 20. Determine a primitive function of the function $f(x) = \dfrac{x-1}{x\left(\sqrt{x}+\sqrt[3]{x^2}\right)}$.

Solution. The function is continuous on $D_f = (0,+\infty)$ and thus has a primitive function there.

If there are expressions

$$\left(\frac{ax+b}{cx+d}\right)^{\frac{p_1}{q_1}}, \ldots, \left(\frac{ax+b}{cx+d}\right)^{\frac{p_n}{q_n}},$$

where $a,b,c,d \in \mathbb{R}$, $ad-bc \neq 0$, $p_1,\ldots,p_n \in \mathbb{Z}$, $q_1,\ldots,q_n \in \mathbb{N}$ in the function formula of f, then we use the substitution $t = \left(\frac{ax+b}{cx+d}\right)^{\frac{1}{s}}$, where s is the least common multiple of the numbers $q_1,\ldots,q_n$.

In our case we have the powers $x^{1/2}$ and $x^{2/3}$ in the function formula of f. The least common multiple of numbers 2 and 3 is 6. We thus use the substitution $t = x^{1/6}$, $x \in (0,+\infty)$. Hence we could derive $\mathrm{d}x = 6t^5\,\mathrm{d}t$. Then we are searching for a primitive function over the interval $(0,+\infty)$

$$\int \frac{t^6-1}{t^6(t^3+t^4)} \cdot 6t^5\,\mathrm{d}t = 6\int \frac{t^6-1}{t^5+t^4}\,\mathrm{d}t.$$

Since in the last integrand (it is a rational function of the variable t) is the degree of the numerator greater then the degree of the denominator, we have to divide first:

$$\begin{aligned}(t^6-1):(t^5+t^4) &= t-1+\frac{t^4-1}{t^4(t+1)}= \\ &= t-1+\frac{(t-1)(t+1)(t^2+1)}{t^4(t+1)}= \\ &= t-1+\frac{1}{t}-\frac{1}{t^2}+\frac{1}{t^3}-\frac{1}{t^4}.\end{aligned}$$

Now we could integrate

$$6\int\frac{t^6-1}{t^5+t^4}\,\mathrm{d}t \overset{c}{=} 3t^2-6t+6\log t+6\frac{1}{t}-3\frac{1}{t^2}+2\frac{1}{t^3},\quad t\in(0,+\infty).$$

From the theorem about integration by substitution, a primitive function of f over $(0,+\infty)$ is every function of the form

$$3\sqrt[3]{x}-6\sqrt[6]{x}+\log x+6\frac{1}{\sqrt[6]{x}}-3\frac{1}{\sqrt[3]{x}}+2\frac{1}{\sqrt{x}}+c,$$

where $c\in\mathbb{R}$ is an arbitrary constant. ♣

Example 21. Determine a primitive function of the function $f(x)=\dfrac{1}{x+\sqrt{x^2+x+1}}$.

Solution. The function f is continuous on the domain $D_f=(-\infty,-1)\cup(-1,+\infty)$. The expression under a radical sign is positive on the whole $\mathbb{R}$, we thus use the Euler substitution $\sqrt{x^2+x+1}=x+t$. By exponentiating we obtain $x^2+x+1=x^2+2xt+t^2$, i.e. $x=\frac{t^2-1}{1-2t}$, and calculate $\mathrm{d}x=-2\frac{t^2-t+1}{(1-2t)^2}\,\mathrm{d}t$. We need to express the formula $\sqrt{x^2+x+1}$ in terms of a new variable t, which is simple:

$$\sqrt{x^2+x+1}=x+t=\frac{t^2-1}{1-2t}+t.$$

Now we substitute and after manipulation we obtain

$$\int\frac{2t^2-2t+2}{(t-2)(2t-1)}\,\mathrm{d}t.$$

Let us realize that we use the Theorem 4(ii) for $\varphi(t)=\frac{t^2-1}{1-2t}$. Next it follows that $\varphi'(t)=-2\frac{t^2-t+1}{(1-2t)^2}<0$ for $t\in\left(-\infty,\frac{1}{2}\right)\cup\left(\frac{1}{2},+\infty\right)$, $\varphi\left(\left(\frac{1}{2},2\right)\right)=(-1,+\infty)$ and $\varphi\big((2,+\infty)\big)=(-\infty,-1)$.

The achieved rational function has the degree of the polynomial in the numerator the same as the degree of the polynomial in the denominator, thus we have to

divide first:

$$(2t^2-2t+2):(2t^2-5t+2)=1+\frac{3t}{(t-2)(2t-1)}.$$

We decompose the second summand to partial fractions and obtain

$$\int\frac{2t^2-2t+2}{(t-2)(2t-1)}\,\mathrm{d}t=\int 1\,\mathrm{d}t+2\int\frac{1}{t-2}-\int\frac{1}{2t-1}\,\mathrm{d}t\overset{c}{=}$$
$$\overset{c}{=}t+2\log|t-2|-\frac{1}{2}\log|2t-1|$$

on the intervals $\left(\frac{1}{2},2\right)$ and $(2,+\infty)$.

According to the Theorem 4(ii) the primitive function of the function f over each of the intervals $(-\infty,-1)$ and $(-1,+\infty)$ is of the form

$$\sqrt{x^2+x+1}-x+2\log\left|\sqrt{x^2+x+1}-x-2\right|-$$
$$-\frac{1}{2}\log\left|2\sqrt{x^2+x+1}-2x-1\right|+c,\quad c\in\mathbb{R}.\quad\clubsuit$$

3.2. Riemann integral

The introduction of the Riemann integral is motivated, among other things, by a problem how to define an area of more complicated sets in a plane than are basic geometric shapes like a rectangle, a triangle etc.

Let f be a bounded non-negative function defined on a bounded closed interval $[a,b]$. We want to define an area under the graph of the function f to be consistent with measuring an area of basic geometric shapes.

One of the possibilities is to aproximate the shape by finite unions of rectangles with known areas and then "in the limit" get the area of the shape. The whole idea is illustrated on the following figures. At the first two figures there are upper approximations of the area of the shape, on the second two there are lower approximations on the contrary.

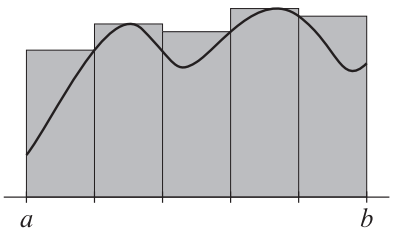


FIGURE 3.

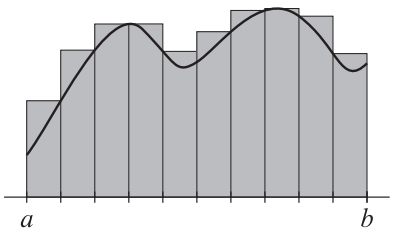


FIGURE 4.

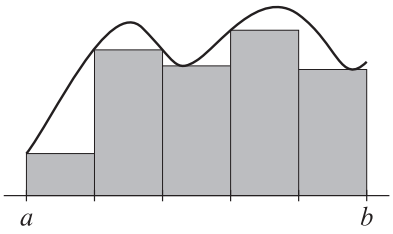


FIGURE 5.

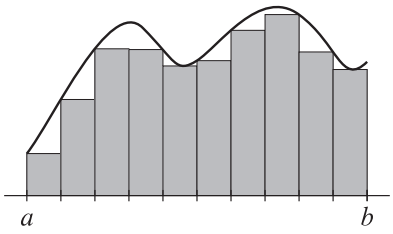


FIGURE 6.

Let us express this intuitive reasoning by exact mathematical notions.

Definition. Let $a, b \in \mathbb{R}$, $a < b$. A finite sequence $\{x_j\}_{j=0}^n$ is the **partition of the interval** $[a, b]$ provided that

$$a = x_0 < x_1 < \cdots < x_n = b.$$

We call the points $x_0, \ldots, x_n$ the **parting points**. We say that partition D' of the interval $[a, b]$ is the **refinement of the partition** D of the interval $[a, b]$, if each parting point of D is also a parting point of D'.

Definition. Let $a, b \in \mathbb{R}$, $a < b$, the function f be bounded on the interval $[a, b]$ and $D = \{x_j\}_{j=0}^n$ be a partition of $[a, b]$. Let us put

$$\overline{S}(f, D) = \sum_{j=1}^{n} M_j(x_j - x_{j-1}), \text{ where } M_j = \sup_{[x_{j-1}, x_j]} f,$$

$$\underline{S}(f, D) = \sum_{j=1}^{n} m_j(x_j - x_{j-1}), \text{ where } m_j = \inf_{[x_{j-1}, x_j]} f,$$

$$\overline{\int_a^b} f = \inf\big\{\overline{S}(f, D);\ D \text{ is a partition of the interval } [a, b]\big\},$$

$$\underline{\int_a^b} f = \sup\big\{\underline{S}(f, D);\ D \text{ is a partition of the interval } [a, b]\big\}.$$

We say that the function f has the **Riemann integral** over the interval $[a, b]$ if $\overline{\int_a^b} f = \underline{\int_a^b} f$. The value of the integral of the function f over the interval $[a, b]$ is then equal to the identical value of $\overline{\int_a^b} f$ and $\underline{\int_a^b} f$ and we denote it by $\int_a^b f$. If $a > b$, we define $\int_a^b f = -\int_b^a f$, and in the case that $a = b$, we define $\int_a^b f = 0$.

Remarks. 1. We call the number $\overline{S}(f, D)$ ($\underline{S}(f, D)$) the **upper** (**lower**, respectively) **sum** with partition D and we call $\overline{\int_a^b} f$ ($\underline{\int_a^b} f$) the **upper** (**lower**, respectively) **integral**.

2. From the boundedness of f on the interval $[a, b]$ it follows that $\overline{S}(f, D) \in \mathbb{R}$ and $\underline{S}(f, D) \in \mathbb{R}$ for each partition D of the interval $[a, b]$ and also both $\overline{\int_a^b} f \in \mathbb{R}$ and $\underline{\int_a^b} f \in \mathbb{R}$. This follows i.e. from the following inequalities:

$$(b - a) \inf_{[a,b]} f \le \overline{S}(f, D) \le (b - a) \sup_{[a,b]} f.$$

3. Traditionally the symbols $\int_a^b f(x)\,\mathrm{d}x$, $\int_a^b f(t)\,\mathrm{d}t$, etc. could be used for the Riemann integral instead of $\int_a^b f$, especially in cases where the variable of the function f need to be emphasized.

Remark. Let f be a bounded non-negative function on an interval $[a,b]$. If the function f has the Riemann integral over $[a,b]$, then the number $\int_a^b f$ can be taken as the area of the set under the graph of the function f, i.e. the set $\{[x,y] \in \mathbb{R}^2;\ a \leq x \leq b, 0 \leq y \leq f(x)\}$.

Example 22. It can be shown easily from definition that for the function $f(x) = 1$, $x \in [0,1]$ follows $\int_0^1 f = 1$, since in this case $\overline{S}(f,D) = \underline{S}(f,D) = 1$ holds for each partition D of the interval $[0,1]$. Similarly $\int_a^b c\,\mathrm{d}x = c(b-a)$ holds for each constant function $x \mapsto c \in \mathbb{R}$ a $a, b \in \mathbb{R}$.

Example 23. Let a function $f\colon \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 1 & \text{for } x \in \mathbb{Q}, \\ 0 & \text{for } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

We call it the **Dirichlet** function. From the Theorem **??** it follows that $\overline{S}(f,D) = 1$ and $\underline{S}(f,D) = 0$ for each partition D of the interval $[0,1]$. Thus, $\overline{\int_0^1} f = 1$ and $\underline{\int_0^1} f = 0$ hold and therefore the Dirichlet function has not the Riemann integral over the interval $[0,1]$.

From the foregoing example it could be seen that not every bounded function has the Riemann integral. In this section we will derive some properties of the Riemann integral and we will show that at least the continuous functions on a bounded interval have the Riemann integral. In the following remark we will look first at some properties of the upper and lower sums.

Remark. Let $a, b \in \mathbb{R}$, $a < b$, and $f\colon [a,b] \to \mathbb{R}$ be a bounded function. From the definition it can be seen at once that for an arbitrary partition D of the interval $[a,b]$ we have

$$\underline{S}(f,D) \leq \overline{S}(f,D). \tag{5}$$

Let now D, D' be partititons of the interval $[a,b]$ and D' be the refinement of D. Then it is not difficult to show that

$$\underline{S}(f,D) \leq \underline{S}(f,D') \leq \overline{S}(f,D') \leq \overline{S}(f,D). \tag{6}$$

Let D_1, D_2 be two arbitrary partitions of the interval $[a,b]$. Let D' be a refinement of both D_1 and D_2. (It suffices to take for D' a partition which consist of all points from both D_1 and D_2.) According to (6) it then follows

$$\underline{S}(f,D_1) \leq \underline{S}(f,D') \leq \overline{S}(f,D') \leq \overline{S}(f,D_2). \tag{7}$$

Hence, it could be easily derived that the following inequality always holds

$$\underline{\int_a^b} f \leq \overline{\int_a^b} f. \tag{8}$$

In fact, according to the (7) $\underline{\int_a^b} f \leq \overline{S}(f, D_2)$ holds for a fixed partition D_2. The inequality (8) now follows from the definition of an infimum.

If we summarize the foregoing thoughts, we obtain that for arbitrary two partitions D_1 and D_2 of the interval $[a, b]$ follows

$$\underline{S}(f, D_1) \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq \overline{S}(f, D_2). \tag{9}$$

Now we will prove a key lemma which enables us to avoid the notions of upper and lower integral by using the Riemann integral.

Lemma 24 (criterion of existence of Riemann integral)**.** Let $a, b \in \mathbb{R}$, $a < b$, and f be a bounded function on the interval $[a, b]$.

(i) $\int_a^b f = I \in \mathbb{R}$ if and only if for each $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$ there exists a partition D of the interval $[a, b]$ such that

$$I - \varepsilon < \underline{S}(f, D) \leq \overline{S}(f, D) < I + \varepsilon. \tag{10}$$

(ii) The function f has the Riemann integral over $[a, b]$ if and only if for each $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$ there exists a partition D of the interval $[a, b]$ such that

$$\overline{S}(f, D) - \underline{S}(f, D) < \varepsilon. \tag{11}$$

Proof. Let us first prove the assertion (i).

$\Rightarrow$ Let us choose an arbitrary $\varepsilon > 0$. Since $\underline{\int_a^b} f = I$ there exists a partition D_1 of the interval $[a, b]$ such that $\underline{S}(f, D_1) > I - \varepsilon$. Similarly $\overline{\int_a^b} f = I$ holds and thus there exists a partition D_2 of the interval $[a, b]$ such that $\overline{S}(f, D_2) < I + \varepsilon$. Let D is a refinement of both D_1 and D_2 on the interval $[a, b]$. Then according to the (7) follows

$$I - \varepsilon < \underline{S}(f, D_1) \leq \underline{S}(f, D) \leq \overline{S}(f, D) \leq \overline{S}(f, D_2) < I + \varepsilon.$$

$\Leftarrow$ Let us choose an arbitrary $\varepsilon > 0$. Let us find a partition D of the interval $[a, b]$ which satisfies the inequalities (10) for the given ε. From the inequalities (9) and (10) follows

$$I - \varepsilon < \underline{S}(f, D) \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq \overline{S}(f, D) < I + \varepsilon.$$

We thus have for each $\varepsilon > 0$

$$I - \varepsilon < \underline{\int_a^b} f \leq \overline{\int_a^b} f < I + \varepsilon,$$

which means that $\underline{\int_a^b} f = \overline{\int_a^b} f = I$.

Now we prove the assertion (ii).

$\Rightarrow$ This implication follows from assertion (i), which was proved above.

$\Leftarrow$ Let us choose an arbitrary $\varepsilon > 0$. Let us find a partition D of the interval $[a, b]$ which satisfies the inequality (11) for the given ε. From (9) and (11) we thus obtain

$$0 \leq \overline{\int_a^b} f - \underline{\int_a^b} f \leq \overline{S}(f, D) - \underline{S}(f, D) < \varepsilon.$$

Since ε is an arbitrary positive number, it must be $\underline{\int_a^b} f = \overline{\int_a^b} f$. ∎

In the following assertions we will show some basic properties of the Riemann integral.

Theorem 25.

(i) Let a function f have the Riemann integral over the interval $[a, b]$ and let $[c, d] \subset [a, b]$. Then f has the Riemann integral also over the interval $[c, d]$.

(ii) Let $c \in (a, b)$ and a function f have the Riemann integral over the intervals $[a, c]$ and $[c, b]$. Then f has the Riemann integral over $[a, b]$ and the following equality holds

$$\int_a^b f = \int_a^c f + \int_c^b f. \tag{12}$$

Proof. (i) Let us prove the assertion for the case $a < b < c < d$, the other cases can be proved similarly. Let us choose an arbitrary $\varepsilon > 0$. According to the Lemma 24(ii) there exists a partition D of the interval $[a, b]$ satisfying the inequality (11). According to (6) we could assume without loss of generality that the partition D contains both the points c and d. Let a partition of the interval $[a, c]$, which contains all parting points of D from the interval $[a, c]$, be denoted by D_1, a partition of the interval $[c, d]$, which contains all parting points of D from the interval $[c, d]$, be denoted by D_2 and a partition of the interval $[d, b]$, which contains all parting points of D from the interval $[d, b]$, be denoted by D_3.

It follows obviously that

$$\overline{S}(f, D) = \overline{S}(f, D_1) + \overline{S}(f, D_2) + \overline{S}(f, D_3),$$
$$\underline{S}(f, D) = \underline{S}(f, D_1) + \underline{S}(f, D_2) + \underline{S}(f, D_3).$$

Applying (5) we thus get

$$\begin{aligned} 0 &\leq \overline{S}(f, D_2) - \underline{S}(f, D_2) \leq \\ &\leq \overline{S}(f, D_2) - \underline{S}(f, D_2) + \overline{S}(f, D_1) - \underline{S}(f, D_1) + \overline{S}(f, D_3) - \underline{S}(f, D_3) = \\ &= \overline{S}(f, D) - \underline{S}(f, D) < \varepsilon. \end{aligned}$$

According to Lemma 24(ii), we thus obtain that the Riemann integral $\int_c^d f$ exists.

(ii) Let us put $I_1 = \int_a^c f$ and $I_2 = \int_c^b f$. Let us choose an arbitrary $\varepsilon > 0$. According to Lemma 24(i) there exists a partition D_1 of the interval $[a, c]$ and a partition D_2 of the interval $[c, b]$ satisfying

$$I_1 - \frac{\varepsilon}{2} < \underline{S}(f, D_1) \leq \overline{S}(f, D_1) < I_1 + \frac{\varepsilon}{2}, \tag{13}$$

$$I_2 - \frac{\varepsilon}{2} < \underline{S}(f, D_2) \leq \overline{S}(f, D_2) < I_2 + \frac{\varepsilon}{2}. \tag{14}$$

Let D be a partition of the interval $[a, b]$, which consist of all the points of the partition D_1 and the partition D_2. Then obviously

$$\overline{S}(f, D) = \overline{S}(f, D_1) + \overline{S}(f, D_2) \quad \text{a} \quad \underline{S}(f, D) = \underline{S}(f, D_1) + \underline{S}(f, D_2).$$

By adding the inequalities (13) and (14) together we obtain

$$I_1 + I_2 - \varepsilon < \underline{S}(f, D) \leq \overline{S}(f, D) < I_1 + I_2 + \varepsilon.$$

According to Lemma 24(i) is thus $\int_a^b f = I_1 + I_2$. ∎

Remark. It can be easily realized that the formula (12) holds for every $a, b, c \in \mathbb{R}$, provided that there exists an integral of the function f over the interval $\big[\min\{a, b, c\}, \max\{a, b, c\}\big]$.

Theorem 26 (linearity of Riemann integral)**.** Let f and g be functions which have the Riemann integral over the interval $[a, b]$ and let $\alpha \in \mathbb{R}$. Then

(i) the function αf has the Riemann integral over $[a, b]$ and

$$\int_a^b \alpha f = \alpha \int_a^b f,$$

(ii) the function $f + g$ has the Riemann integral over $[a, b]$ and

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g. \tag{15}$$

Proof. Let us notice that for $a = b$ are both assertions obvious. In what follows, we thus suppose that $a < b$.

(i) Assume first that $\alpha \geq 0$. Then for each partition D of the interval $[a, b]$ follows $\overline{S}(\alpha f, D) = \alpha \overline{S}(f, D)$ and $\underline{S}(\alpha f, D) = \alpha \underline{S}(f, D)$. Thus we obtain at once $\overline{\int_a^b} \alpha f = \alpha \overline{\int_a^b} f$ and $\underline{\int_a^b} \alpha f = \alpha \underline{\int_a^b} f$, which gives us the required equality.

Next, let $D = \{x_i\}_{i=0}^n$ be an arbitrary partition of the interval $[a, b]$. Since $\sup_{[x_{i-1}, x_i]}(-f) = -\inf_{[x_{i-1}, x_i]} f$ holds for $i = 1, \ldots, n$ (see the remark on the

page **??**), we obtain $\overline{S}(-f,D) = -\underline{S}(f,D)$. Thus we have

$$\begin{aligned}\overline{\int_a^b}(-f) &= \inf\big\{\overline{S}(-f,D);\ D \text{ is a partition of } [a,b]\big\} = \\ &= \inf\big\{-\underline{S}(f,D);\ D \text{ is a partition of } [a,b]\big\} = \\ &= -\sup\big\{\underline{S}(f,D);\ D \text{ is a partition of } [a,b]\big\} = \\ &= -\underline{\int_a^b} f = -\int_a^b f.\end{aligned}$$

we could obtain similarly $\underline{\int_a^b}(-f) = -\int_a^b f$.

Finally according to the previous we get for $\alpha < 0$

$$\int_a^b \alpha f = \int_a^b \big(-|\alpha|\, f\big) = -\int_a^b |\alpha|\, f = -|\alpha| \int_a^b f = \alpha \int_a^b f.$$

(ii) Let us choose an arbitrary $\varepsilon > 0$. From the Lemma 24(i) we can find partitions D_1 and D_2 of the interval $[a,b]$ such that

$$\begin{aligned}\int_a^b f - \frac{\varepsilon}{2} < \underline{S}(f,D_1) \le \overline{S}(f,D_1) < \int_a^b f + \frac{\varepsilon}{2},\\ \int_a^b g - \frac{\varepsilon}{2} < \underline{S}(g,D_2) \le \overline{S}(g,D_2) < \int_a^b g + \frac{\varepsilon}{2}.\end{aligned}$$

Let D be a refinement of both D_1 and D_2 on the interval $[a,b]$. Then according to (7) it follows that

$$\begin{aligned}\int_a^b f - \frac{\varepsilon}{2} < \underline{S}(f,D) \le \overline{S}(f,D) < \int_a^b f + \frac{\varepsilon}{2},\\ \int_a^b g - \frac{\varepsilon}{2} < \underline{S}(g,D) \le \overline{S}(g,D) < \int_a^b g + \frac{\varepsilon}{2}.\end{aligned} \tag{16}$$

From the definitions of the upper and lower sums and from the Example **??**, we obtain the inequalities $\overline{S}(f+g,D) \le \overline{S}(f,D) + \overline{S}(g,D)$ and $\underline{S}(f+g,D) \ge \underline{S}(f,D) + \underline{S}(g,D)$. This together with the inequalities (16) yields

$$\begin{aligned}\int_a^b f + \int_a^b g - \varepsilon < \underline{S}(f,D) + \underline{S}(g,D) \le \underline{S}(f+g,D) \le \\ \le \overline{S}(f+g,D) \le \overline{S}(f,D) + \overline{S}(g,D) < \int_a^b f + \int_a^b g + \varepsilon.\end{aligned}$$

According to Lemma 24(i) the equality (15) thus holds. ∎

Theorem 27. Let $a, b \in \mathbb{R}$, $a < b$, and let f and g be functions which have the Riemann integral over the interval $[a,b]$. Then:

(i) If $f(x) \leq g(x)$ for each $x \in [a,b]$, then $\int_a^b f \leq \int_a^b g$.

(ii) The function $|f|$ has the Riemann integral over $[a,b]$ and $\left|\int_a^b f\right| \leq \int_a^b |f|$ holds.

Proof. (i) Applying Theorem 26, we obtain $\int_a^b g - \int_a^b f = \int_a^b (g-f) \geq 0$, since $\underline{S}(g-f,D) \geq 0$ holds for each partition D of the interval $[a,b]$.

(ii) If we know that the function $|f|$ has the Riemann integral over $[a,b]$, then the required equality could be proved easily. It is because $-|f(t)| \leq f(t) \leq |f(t)|$ holds for each $t \in \langle a,b\rangle$ and thus according to (i) and Theorem 26(i) we have

$$-\int_a^b |f| = \int_a^b -|f| \leq \int_a^b f \leq \int_a^b |f|\,.$$

From that we obtain the equality from assertion and the proof is completed.

Let us prove now that the function $|f|$ has the Riemann intergral over $[a,b]$. Let us choose $\varepsilon > 0$, then from the Lemma 24(ii) we could find a partition $D = \{x_i\}_{i=0}^n$ of the interval $[a,b]$ such that $\overline{S}(f,D) - \underline{S}(f,D) < \varepsilon$. For $i = 1,\ldots,n$ let us put

$$M_i = \sup_{[x_{i-1},x_i]} f, \quad m_i = \inf_{[x_{i-1},x_i]} f,$$
$$\widehat{M}_i = \sup_{[x_{i-1},x_i]} |f|\,, \quad \widehat{m}_i = \inf_{[x_{i-1},x_i]} |f|\,.$$

According to Example **??** the inequality $\widehat{M}_i - \widehat{m}_i \leq M_i - m_i$ holds for $i = 1,\ldots,n$. By using these inequalities we obtain

$$\begin{aligned}
\overline{S}(|f|\,,D) - \underline{S}(|f|\,,D) &= \sum_{i=1}^n \widehat{M}_i(x_i - x_{i-1}) - \sum_{i=1}^n \widehat{m}_i(x_i - x_{i-1}) = \\
&= \sum_{i=1}^n (\widehat{M}_i - \widehat{m}_i)(x_i - x_{i-1}) \leq \\
&\leq \sum_{i=1}^n (M_i - m_i)(x_i - x_{i-1}) = \\
&= \sum_{i=1}^n M_i(x_i - x_{i-1}) - \sum_{i=1}^n m_i(x_i - x_{i-1}) = \\
&= \overline{S}(f,D) - \underline{S}(f,D) < \varepsilon.
\end{aligned}$$

However, according to Lemma 24(ii) it means that the integral $\int_a^b |f|$ exists. ∎

To show that every continuous function on a closed interval has the Riemann integral, the following notion will be needed.

Definition. We say that a function f is **uniformly continuous on an interval** I, provided that

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \exists \delta \in \mathbb{R}, \delta > 0\ \forall x, y \in I, |x-y| < \delta\colon\ |f(x) - f(y)| < \varepsilon.$$

Remark. Let us note the difference between the definition of a function which is continuous on the interval I and the definition of a function which is uniformly continuous on the interval I. The function f is continuous on I if and only if it is continuous at each point of I with respect to I, in other words

$$\forall x \in I\ \forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \exists \delta \in \mathbb{R}, \delta > 0\ \forall y \in I, |x-y| < \delta\colon\ |f(x) - f(y)| < \varepsilon,$$

which is equivalent to

$$\forall \varepsilon \in \mathbb{R}, \varepsilon > 0\ \forall x \in I\ \exists \delta \in \mathbb{R}, \delta > 0\ \forall y \in I, |x-y| < \delta\colon\ |f(x) - f(y)| < \varepsilon.$$

Both definitions thus differ only in the order of quantifiers. The difference is that in definition of continuity for a given ε we are searching for δ separately for each point $x \in I$ (and thus the value δ generally depends on x and can be different for different x), in definition of uniform continuity for a given ε we are searching for one δ, and this δ-neighbourhood is then used at each point $x \in I$, in other words this δ is the same for all $x \in I$.

From what we have just said it can be seen that every uniformly continuous function on I is a continuous function on I. However, the converse implication generally does not hold. It is not difficult to show that the function $f(x) = 1/x$, $x \in (0,1)$, is continuous on this interval, but not uniformly continuous.

Theorem 28. If a function f is continuous on a bounded closed interval $[a,b]$, then it is uniformly continuous on $[a,b]$.

Proof. Let us suppose that f is continuous, but not uniformly continuous on $[a,b]$. Then there exists $\varepsilon > 0$ such that

$$\forall \delta \in \mathbb{R}, \delta > 0\ \exists x, y \in [a,b], |x-y| < \delta\colon\ |f(x) - f(y)| \geq \varepsilon.$$

Hence we get especially that for each $n \in \mathbb{N}$ there exist $x_n, y_n \in [a,b]$ satisfying $|x_n - y_n| < 1/n$ and $|f(x_n) - f(y_n)| \geq \varepsilon$. Due to compactness of the set $[a,b]$, we could from the sequence $\{x_n\}_{n=1}^{\infty}$ choose a convergent subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ which converges to $x \in [a,b]$. Concurrently it must hold $\lim\limits_{k\to\infty} y_{n_k} = x$, since

$$|y_{n_k} - x| \leq |y_{n_k} - x_{n_k}| + |x_{n_k} - x| \leq 1/n_k + |x_{n_k} - x|\,.$$

The function f is continuous at x, hence according to the Heine theorem follows $\lim\limits_{k\to\infty} f(x_{n_k}) = f(x)$ and $\lim\limits_{k\to\infty} f(y_{n_k}) = f(x)$. On the other hand we have $|f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon$ and that is a contradiction. ∎

Theorem 29. Let the function f be continuous on the interval $[a,b]$. Then f has the Riemann integral on $[a,b]$.

Proof. Let us choose an arbitrary $\varepsilon > 0$. According to the previous theorem we could find $\delta > 0$ such that

$$\forall x, y \in [a,b], |x-y| < \delta \colon\ |f(x) - f(y)| < \frac{\varepsilon}{b-a}.$$

Let us choose a partition $D = \{x_j\}_{j=0}^n$ such that $x_j - x_{j-1} < \delta$, $j = 1, \ldots, n$. From the Example **??** we obtain

$$M_j - m_j = \sup_{[x_{j-1}, x_j]} f - \inf_{[x_{j-1}, x_j]} f \leq \frac{\varepsilon}{b-a}.$$

It thus follows that

$$\begin{aligned} 0 \leq \overline{S}(f,D) - \underline{S}(f,D) &= \sum_{j=1}^n (M_j - m_j)(x_j - x_{j-1}) \leq \\ &\leq \sum_{j=1}^n \frac{\varepsilon}{b-a}(x_j - x_{j-1}) = \\ &= \frac{\varepsilon}{b-a}(b-a) = \varepsilon. \end{aligned}$$

Now we use Lemma 24(ii) once more. ∎

Theorem 30. Let f be a continuous function on an interval (a,b) and let $c \in (a,b)$. If we put $F(x) = \int_c^x f$ for $x \in (a,b)$, then $F'(x) = f(x)$ for each $x \in (a,b)$, in other words F is a primitive function of f over (a,b).

Proof. According to Theorem 29 $F(x) \in \mathbb{R}$ holds for each $x \in (a,b)$, so F is a real function defined on (a,b). Let us now choose a point $x \in (a,b)$ fixedly. We want to show that $F'(x) = f(x)$ holds, in other words

$$\lim_{h \to 0} \left(\frac{F(x+h) - F(x)}{h} - f(x) \right) = 0.$$

Let us choose an arbitrary $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$. Since the function f is continuous at the point x, we could find $\delta \in \mathbb{R}$, $0 < \delta \leq \min\{b-x, x-a\}$, such that for each $t \in B(x,\delta)$ the inequality $|f(t) - f(x)| < \varepsilon$ holds. Let us take now $h \in P(0,\delta)$.

Then we have $x+h \in (a,b)$. For $h>0$ we could write

$$\begin{aligned}
&\left|\frac{1}{h}\big(F(x+h)-F(x)\big)-f(x)\right| = \\
&\qquad = \left|\frac{1}{h}\left(\int_c^{x+h} f(t)\,\mathrm{d}t - \int_c^{x} f(t)\,\mathrm{d}t\right) - f(x)\right| = \\
&\qquad = \left|\frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t - f(x)\right| = \\
&\qquad = \left|\frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t - f(x)\cdot\frac{1}{h}\int_x^{x+h} 1\,\mathrm{d}t\right| = \\
&\qquad = \frac{1}{h}\cdot\left|\int_x^{x+h}\big(f(t)-f(x)\big)\,\mathrm{d}t\right| \le \\
&\qquad \le \frac{1}{h}\int_x^{x+h} |f(t)-f(x)|\,\mathrm{d}t \le \\
&\qquad \le \frac{1}{h}\int_x^{x+h} \varepsilon\,\mathrm{d}t = \frac{1}{h}\cdot h\varepsilon = \varepsilon.
\end{aligned}$$

(We used in this order Theorem 25(ii) together with the following remark, Theorem 26, Theorem 27(ii) and Theorem 27(i).) For $h<0$ can be shown the same inequality similarly, it is only necessary to pay attention to the fact that in this case we have $x+h<x$. Hence, it is justified $F'(x)=\lim\limits_{h\to 0}\frac{1}{h}\big(F(x+h)-F(x)\big)=f(x)$. ∎

The previous theorem enables us to prove Theorem 3 about the existence of a primitive function.

The proof of Theorem 3. Let us choose $c\in(a,b)$ and put

$$F(x)=\int_c^x f(t)\,\mathrm{d}t, \quad x\in(a,b).$$

The function F is defined on the whole interval (a,b) (Theorem 29) and according to Theorem 30 for each $x\in(a,b)$ $F'(x)=f(x)$ holds. The function F is thus a primitive function of f over (a,b). ∎

The following theorem gives instructions, how to calculate the Riemann integral from a primitive function.

Theorem 31 (Newton-Leibniz formula)**.** Let f be continuous on a bounded closed interval $[a,b]$, $a<b$, and F is a primitive function of f over (a,b). Then there exist limits $\lim\limits_{x\to a+} F(x)\in\mathbb{R}$, $\lim\limits_{x\to b-} F(x)\in\mathbb{R}$ and

$$\int_a^b f = \lim_{x\to b-} F(x) - \lim_{x\to a+} F(x). \tag{17}$$

Proof. Let the function $\tilde{f}$ be defined on the interval $[a-1,b+1]$ by:

$$\tilde{f}(x) = \begin{cases} f(a) & \text{for } x\in(a-1,a),\\ f(x) & \text{for } x\in[a,b],\\ f(b) & \text{for } x\in(b,b+1). \end{cases}$$

Let next $G\colon (a-1,b+1)\to\mathbb{R}$ be defined by $G(x)=\int_a^x \tilde{f}$. The function $\tilde{f}$ is continuous on $(a-1,b+1)$, according to Theorem 30 is thus G a primitive function of $\tilde{f}$ over $(a-1,b+1)$. The function $G|_{(a,b)}$ is a primitive function of f over (a,b) and therefore there exists $c\in\mathbb{R}$ such that $F=G|_{(a,b)}+c$. The function G is continuous at the points a and b, hence there exists limits $\lim\limits_{x\to b-} F(x) = \lim\limits_{x\to b-} G(x)+c\in\mathbb{R}$, $\lim\limits_{x\to a+} F(x) = \lim\limits_{x\to a+} G(x)+c\in\mathbb{R}$. We thus have

$$\begin{aligned}\int_a^b f = G(b) = G(b)-G(a) &= \lim_{x\to b-} G(x) - \lim_{x\to a+} G(x) = \\ &= \Big(\lim_{x\to b-} F(x) - c\Big) - \Big(\lim_{x\to a+} F(x) - c\Big) = \\ &= \lim_{x\to b-} F(x) - \lim_{x\to a+} F(x).\end{aligned}$$

■

Remark. Let us put

$$[F]_a^b = \begin{cases} \lim\limits_{x\to b-} F(x) - \lim\limits_{x\to a+} F(x) & \text{for } a<b,\\ \lim\limits_{x\to b+} F(x) - \lim\limits_{x\to a-} F(x) & \text{for } b<a. \end{cases}$$

Then the Newton-Leibniz formula can be written as

$$\int_a^b f = [F]_a^b$$

also for $b<a$.

From the Newton-Leibniz formula follows the following two theorems often used in calculation.

Theorem 32 (integration by parts)**.** Let the functions f, g, f' and g' be continuous on an interval $[a,b]$.[1] Then

$$\int_a^b f'g = [fg]_a^b - \int_a^b fg'.$$

Proof. The function fg is a primitive function of the function $f'g + fg'$ over the interval (a,b). Therefore

$$\int_a^b (f'g + fg') = [fg]_a^b.$$

holds according Theorem 31. The formula then follows from Theorem 26. ∎

Theorem 33 (integration by substitution)**.** Let the function f be continuous on an interval $[a,b]$. Let next the function φ have a continuous derivative on an interval $[\alpha,\beta]$ and map it to the interval $[a,b]$. Then

$$\int_\alpha^\beta f\big(\varphi(x)\big)\varphi'(x)\,\mathrm{d}x = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(t)\,\mathrm{d}t.$$

Proof. Let us notice first that due to continuity both of the integrals exist. Let the function $\tilde{f}$ be defined on the interval $(a-1,b+1)$ by:

$$\tilde{f}(t) = \begin{cases} f(a) & \text{for } t \in (a-1,a), \\ f(t) & \text{for } t \in [a,b], \\ f(b) & \text{for } t \in (b,b+1). \end{cases}$$

Let G be a primitive function of $\tilde{f}$ over $(a-1,b+1)$. From the Theorem 31 and Theorem 4(i) (since $\varphi\big((\alpha,\beta)\big) \subset (a-1,b+1)$ holds) we get

$$\int_\alpha^\beta f\big(\varphi(x)\big)\varphi'(x)\,\mathrm{d}x = \int_\alpha^\beta \tilde{f}\big(\varphi(x)\big)\varphi'(x)\,\mathrm{d}x = \big[G\big(\varphi(x)\big)\big]_\alpha^\beta =$$

$$= G\big(\varphi(\beta)\big) - G\big(\varphi(\alpha)\big) = \int_{\varphi(\alpha)}^{\varphi(\beta)} \tilde{f}(t)\,\mathrm{d}t = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(t)\,\mathrm{d}t,$$

and the third equality follows from the continuity of the function $G \circ \varphi$ on the interval $[\alpha,\beta]$. ∎

By using Riemann integral we now prove Theorem **??**.

Theorem. There exist exactly one function $\log$ satisfying these properties:

(i) $D_{\log} = (0,+\infty)$,
(ii) $\log$ is increasing on $(0,+\infty)$,
(iii) $\forall x,y \in (0,+\infty)\colon\ \log xy = \log x + \log y$,

[1] Here, a value of f' at the points a and b stands for corresponding one-sided derivatives.

(iv) $\lim\limits_{x\to 1}\frac{\log x}{x-1}=1$.

Proof. Let us put

$$F(x)=\int_1^x \frac{1}{t}\,\mathrm{d}t,\quad x\in(0,+\infty).$$

We show, that the function F has the required properties.

(i) From Theorem 29 it follows that the function F is defined on the interval $(0,+\infty)$.

(ii) The function F is increasing on the interval $(0,+\infty)$ since $F'(x)=\frac{1}{x}>0$ for each $x\in(0,+\infty)$ according to Theorem 30.

(iii) Let $x>0$ and $y>0$. Then it follows

$$F(xy)=\int_1^{xy}\frac{1}{t}\,\mathrm{d}t=\int_1^x\frac{1}{t}\,\mathrm{d}t+\int_x^{xy}\frac{1}{t}\,\mathrm{d}t=F(x)+\int_x^{xy}\frac{1}{\frac{t}{x}}\cdot\frac{1}{x}\,\mathrm{d}t=$$
$$=F(x)+\int_1^y\frac{1}{z}\,\mathrm{d}z=F(x)+F(y),$$

where the last but one equation follows from Theorem 33 for $\varphi(t)=\frac{t}{x}$.

(iv) Here we have

$$\lim_{x\to 1}\frac{F(x)}{x-1}=\lim_{x\to 1}\frac{F(x)-F(1)}{x-1}=F'(1)=1.$$

Now it remains to prove the uniqueness. Let us suppose that the function G satisfies the conditions of the theorem as well. Then we could derive (similarly to Section **??**) that $G'(x)=\frac{1}{x}$, $x\in(0,+\infty)$, and $G(1)=0$. The function F has also these properties. Therefore, according to Theorem 1, $F=G$ on the interval $(0,+\infty)$ and this is what had to be proved. ∎

Example 34. Evaluate $\displaystyle\int_1^3\frac{1}{x\sqrt{x^2+5x+1}}\,\mathrm{d}x$.

Solution. Let us put $f(x)=\frac{1}{x\sqrt{x^2+5x+1}}$. The function f is continuous on the interval $[1,3]$, it thus has the Riemann integral over this interval. We evaluate it by applying Theorem 31.

We use the Euler substitution $\sqrt{x^2+5x+1}=x+t$ and get

$$x=\frac{1-t^2}{2t-5}\qquad\text{and}\qquad \mathrm{d}x=\frac{-2(t^2-5t+1)}{(2t-5)^2}\,\mathrm{d}t.$$

Then we need to calculate

$$\int\frac{2}{t^2-1}\,\mathrm{d}t\overset{c}{=}\log\frac{t-1}{t+1},\quad t\in(1,+\infty).$$

The function

$$F(x) = \log \frac{\sqrt{x^2+5x+1} - x - 1}{\sqrt{x^2+5x+1} - x + 1}$$

is thus a primitive function of f over the interval $(1, 3)$.

Now it can be calculated easily

$$\int_1^3 f = [F]_1^3 = \log \frac{1}{3} - \log \frac{\sqrt{7}-2}{\sqrt{7}} = \log \frac{\sqrt{7}}{3(\sqrt{7}-2)} = \log \frac{7+2\sqrt{7}}{9}.$$

Another possibility is to use Theorem 33 for calculation. For $\varphi(t) = \frac{1-t^2}{2t-5}$ we have $\varphi\big(\sqrt{7}-1\big) = 1$ and $\varphi(2) = 3$, and thus according to Theorem 33 follows

$$\int_{\sqrt{7}-1}^2 \frac{2}{t^2-1}\,\mathrm{d}t = \int_1^3 \frac{1}{x\sqrt{x^2+5x+1}}\,\mathrm{d}x.$$

This yields

$$\begin{aligned}\int_1^3 \frac{1}{x\sqrt{x^2+5x+1}}\,\mathrm{d}x &= \int_{\sqrt{7}-1}^2 \frac{2}{t^2-1}\,\mathrm{d}t = \\ &= \left[\log \frac{t-1}{t+1}\right]_{\sqrt{7}-1}^2 = \log \frac{7+2\sqrt{7}}{9}.\end{aligned}$$

♣

Example 35. Evaluate $\displaystyle\int_0^\pi \frac{1}{1+3\cos^2 x}\,\mathrm{d}x$.

Solution. According to Example 18

$$F(x) = \begin{cases} \frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right) & \text{for } x \in (0, \frac{\pi}{2}), \\ \frac{\pi}{4} & \text{for } x = \frac{\pi}{2}, \\ \frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right) + \frac{\pi}{2} & \text{for } x \in (\frac{\pi}{2}, \pi) \end{cases}$$

is a primitive function to the integrand over $(0, \pi)$. Then we have

$$\int_0^\pi \frac{1}{1+3\cos^2 x}\,\mathrm{d}x = [F]_0^\pi = 0 + \frac{\pi}{2} - 0 = \frac{\pi}{2}.$$

There is a very frequent mistake in omitting the sticking, i.e. in the wrong reasoning that the function $\frac{1}{2}\operatorname{arctg}\big(\frac{\operatorname{tg} x}{2}\big)$ (which is not defined at the point $\frac{\pi}{2}$) is a primitive function of the integrand over the whole interval $(0, \pi)$. We would thus get

$$\int_0^\pi \frac{1}{1+3\cos^2 x}\,\mathrm{d}x = \left[\frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right)\right]_0^\pi = 0 - 0 = 0.$$

At this point it should surprise us that an integral of a positive continuous function is equal to zero – which is not possible!

Note that at this particular example we could avoid the sticking of a primitive function by using the substitution $t = \operatorname{tg}(x/2)$. Another possibility of solving the problem is using Theorem 25. According to it follows

$$\begin{aligned}\int_0^{\pi} \frac{1}{1+3\cos^2 x}\,\mathrm{d}x &= \int_0^{\frac{\pi}{2}} \frac{1}{1+3\cos^2 x}\,\mathrm{d}x + \int_{\frac{\pi}{2}}^{\pi} \frac{1}{1+3\cos^2 x}\,\mathrm{d}x = \\ &= \left[\frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right)\right]_0^{\frac{\pi}{2}} + \left[\frac{1}{2}\operatorname{arctg}\left(\frac{\operatorname{tg} x}{2}\right)\right]_{\frac{\pi}{2}}^{\pi} = \\ &= \frac{\pi}{4} + \frac{\pi}{4} = \frac{\pi}{2}.\end{aligned}$$

♣

Let us now look at some geometric applications of the definite integral.

Example 36. Let $a, b, p, q \in \mathbb{R}$ and $0 < a < b$, $0 < p < q$. Calculate the area of the shape bounded by graphs of the functions

$$x \mapsto \frac{x^2}{p}, \quad x \mapsto \frac{x^2}{q}, \quad x \mapsto \sqrt{ax} \quad \text{and} \quad x \mapsto \sqrt{bx}.$$

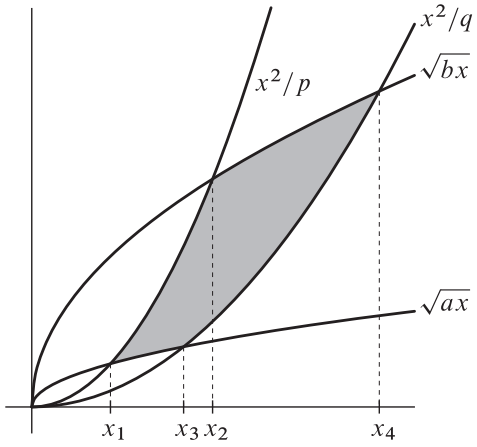


FIGURE 7.

Solution. We could easily calculate x-coordinates of the four intersections of the curves:

$$x_1 = \sqrt[3]{ap^2}, \quad x_2 = \sqrt[3]{bp^2}, \quad x_3 = \sqrt[3]{aq^2} \quad \text{and} \quad x_4 = \sqrt[3]{bq^2}.$$

The area of the shape is equal to

$$\begin{aligned}\int_{x_1}^{x_2} \frac{x^2}{p}\,\mathrm{d}x + \int_{x_2}^{x_4} \sqrt{bx}\,\mathrm{d}x - \int_{x_1}^{x_3} \sqrt{ax}\,\mathrm{d}x - \int_{x_3}^{x_4} \frac{x^2}{q}\,\mathrm{d}x = \\ = \left[\frac{x^3}{3p}\right]_{x_1}^{x_2} + \left[\frac{2\sqrt{bx^3}}{3}\right]_{x_2}^{x_4} - \left[\frac{2\sqrt{ax^3}}{3}\right]_{x_1}^{x_3} - \left[\frac{x^3}{3q}\right]_{x_3}^{x_4} = \\ = \frac{1}{3}(b-a)(q-p).\end{aligned}$$

♣

We could also calculate the length of curves using a definite integral. We will not try to define a curve generally – that is not easy. In our case the curve will be an arbitrary set of the form

$$\{[x,y] \in \mathbb{R}^2;\ a \le x \le b, y = f(x)\},$$

where f is a differentiable function on the interval $[a,b]$, whose derivative is continuous on $[a,b]$.

Let the length of a curve be defined as follows. Let $D = \{x_k\}_{k=0}^n$ is an arbitrary partition of the interval $[a,b]$ and $P_k = [x_k, f(x_k)]$ for $k = 0,1,\ldots,n$. Line segments connecting the points P_{k-1} and P_k, $k = 1,\ldots,n$ form a polygonal chain, whose lenght is

$$\mathrm{l}(D) = \sum_{k=1}^{n} |P_{k-1}P_k|\,,$$

where $|P_{k-1}P_k|$ denotes the length of the line segment, which connects the points P_{k-1} and P_k. The length of a curve is defined to be the number

$$L = \sup\left\{\mathrm{l}(D);\ D \text{ is a partition of the interval } [a,b]\right\}.$$

For the length of the polygonal chain $\mathrm{l}(D)$ it holds:

$$\begin{aligned}\mathrm{l}(D) &= \sum_{k=1}^{n} \sqrt{(x_k - x_{k-1})^2 + \big(f(x_k) - f(x_{k-1})\big)^2} = \\ &= \sum_{k=1}^{n} (x_k - x_{k-1}) \sqrt{1 + \left(\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\right)^2}.\end{aligned}$$

From the Lagrange Mean value theorem (Theorem **??**) we obtain that for each $k \in \{1,\ldots,n\}$ there exists a number $\xi_k \in (x_{k-1}, x_k)$ such that

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} = f'(\xi_k).$$

Hence

$$\mathrm{l}(D) = \sum_{k=1}^{n} \sqrt{1 + \big(f'(\xi_k)\big)^2} \cdot (x_k - x_{k-1}).$$

Let us put $g(x) = \sqrt{1 + \big(f'(x)\big)^2}$. From our assumptions it follows that the function g is continuous on $[a, b]$. Next, we have

$$\sum_{k=1}^{n} \inf_{[x_{k-1}, x_k]} g \cdot (x_k - x_{k-1}) \le \mathrm{l}(D) \le \sum_{k=1}^{n} \sup_{[x_{k-1}, x_k]} g \cdot (x_k - x_{k-1}).$$

The length of the polygonal chain is thus between lower and upper sum of the function g corresponding to the respective partition D. Since g has the Riemann integral over $[a, b]$, it could be deduced that

$$L = \int_a^b g(x)\,\mathrm{d}x = \int_a^b \sqrt{1 + \big(f'(x)\big)^2}\,\mathrm{d}x.$$

By the end we will write (without derivation) formulas for calculation of the surface area and the volume of a solid of revolution. We will take these notions only intuitively and we will not write their exact definitions here.

Let a non-negative continuous function f have a continuous derivative on an interval $[a, b]$. Rotating the graph of this function around the x-axis forms a surface of a solid of revolution, whose area P could be calculated by the formula

$$P = 2\pi \int_a^b f(x) \sqrt{1 + \big(f'(x)\big)^2}\,\mathrm{d}x.$$

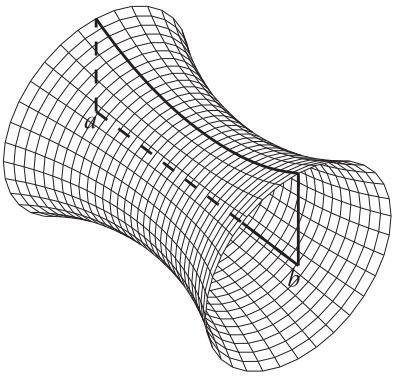


FIGURE 8.

Let f be a continuous non-negative function on an interval $[a, b]$. Then the volume V of a solid of revolution obtained by rotating of the set $\{[x, y] \in \mathbb{R}^2;\ a \le$

$x \le b; 0 \le y \le f(x)\}$ around the x-axis could be calculated by the formula

$$V = \pi \int_a^b f^2(x)\,\mathrm{d}x.$$

Example 37. Calculate the volume of a ball with centre in the origin and the radius $r > 0$.

Solution. Let us put $f(x) = \sqrt{r^2 - x^2}$, $x \in [-r, r]$ and calculate according to the foregoing formula

$$V = \pi \int_{-r}^{r} \left(\sqrt{r^2 - x^2}\right)^2 \mathrm{d}x = \pi \int_{-r}^{r} (r^2 - x^2)\,\mathrm{d}x =$$
$$= \pi \left[r^2 x - \frac{1}{3}x^3\right]_{-r}^{r} = \frac{4}{3}\pi r^3.$$

♣

3.3. Zobecněný Riemannův integrál

V tomto oddílu zobecníme pojem Riemannova integrálu tak, abychom mohli integrovat i některé neomezené funkce a také některé funkce definované na neomezených intervalech.

Lemma 38 (spojitost Riemannova integrálu)**.** Nechť $a, b \in \mathbb{R}$, $a < b$, a funkce f má na intervalu $[a, b]$ Riemannův integrál. Pak platí

$$\int_a^b f = \lim_{x \to b-} \int_a^x f = \lim_{x \to a+} \int_x^b f.$$

Proof. Dokážeme pouze první rovnost, druhou lze dokázat obdobně. Protože f má Riemannův integrál na $[a, b]$, je na $[a, b]$ omezená, a tedy existuje $M > 0$ takové, že $|f(x)| \le M$ pro každé $x \in [a, b]$. Zvolme libovolné $\varepsilon > 0$. Položme $\delta = \min\{\varepsilon/M, b - a\}$. Pak pro $x \in P^-(b, \delta)$ platí

$$\left|\int_a^b f - \int_a^x f\right| = \left|\int_x^b f\right| \le \int_x^b |f| \le \int_x^b M\,\mathrm{d}t = M(b - x) < \varepsilon,$$

přičemž jsme postupně použili Větu 25 a Větu 27. ■

Lemma 39. Nechť $a, b \in \mathbb{R}^*$, $a < b$, a funkce f má Riemannův integrál na každém podintervalu $[x, y] \subset (a, b)$. Nechť dále $c \in (a, b)$, existují limity $\lim_{x \to a+} \int_x^c f$

a $\lim_{y\to b-} \int_c^y f$ a jejich součet má smysl (tj. je definovaný). Pak pro každé $d \in (a,b)$ existují $\lim_{x\to a+} \int_x^d f$ a $\lim_{y\to b-} \int_d^y f$ a platí

$$\lim_{x\to a+} \int_x^d f + \lim_{y\to b-} \int_d^y f = \lim_{x\to a+} \int_x^c f + \lim_{y\to b-} \int_c^y f.$$

Proof. Zvolme libovolné $d \in (a,b)$. Dle předpokladu existuje Riemannův integrál $\int_c^d f$, což je reálné číslo. Platí tedy

$$\begin{aligned}\lim_{x\to a+} \int_x^c f + \lim_{y\to b-} \int_c^y f &= \lim_{x\to a+} \int_x^c f + \lim_{y\to b-} \left(\int_c^d f + \int_d^y f \right) = \\ &= \lim_{x\to a+} \left(\int_x^c f + \int_c^d f \right) + \lim_{y\to b-} \int_d^y f = \\ &= \lim_{x\to a+} \int_x^d f + \lim_{y\to b-} \int_d^y f,\end{aligned}$$

přičemž jsme několikrát použili Větu 25 spolu s poznámkou za ní. ∎

Definition. Nechť $a, b \in \mathbb{R}^*$, $a < b$, a nechť funkce f je definovaná na intervalu (a,b). Má-li funkce f Riemannův integrál na každém podintervalu $[x,y] \subset (a,b)$ a existuje-li $c \in (a,b)$ takové, že limity $\lim_{x\to a+} \int_x^c f$ a $\lim_{y\to b-} \int_c^y f$ existují a jejich součet má smysl, pak definujeme **zobecněný Riemannův integrál** funkce f na intervalu (a,b) jako

$$\int_a^b f = \lim_{x\to a+} \int_x^c f + \lim_{y\to b-} \int_c^y f.$$

Remark. Podle Lemmatu 39 je tato definice korektní, neboť hodnota součtu $\lim_{x\to a+} \int_x^c f + \lim_{y\to b-} \int_c^y f$ nezávisí na volbě dělicího bodu $c \in (a,b)$. Všimněme si, že z Věty 25 a z Lemmatu 38 plyne, že má-li funkce f Riemannův integrál na intervalu $[a,b]$, má i zobecněný Riemannův integrál na intervalu (a,b) a oba integrály jsou si rovny. To nás opravňuje používat symbol $\int_a^b f$ i pro zobecněný Riemannův integrál na intervalu (a,b). Dále si uvědomme, že hodnota *zobecněného* Riemannova integrálu může být i $+\infty$ nebo $-\infty$ na rozdíl od Riemannova integrálu.

Example 40. Zkoumejme existenci následujících zobecněných Riemannových integrálů:

$$\begin{aligned}\int_0^{+\infty} e^{-x}\,\mathrm{d}x &= \lim_{z\to 0+}\int_z^1 e^{-x}\,\mathrm{d}x + \lim_{y\to+\infty}\int_1^y e^{-x}\,\mathrm{d}x = \\ &= \lim_{z\to 0+}\left[-e^{-x}\right]_z^1 + \lim_{y\to+\infty}\left[-e^{-x}\right]_1^y = \\ &= \lim_{z\to 0+}(-e^{-1}+e^{-z}) + \lim_{y\to+\infty}(-e^{-y}+e^{-1}) = 1;\end{aligned}$$

$$\begin{aligned}\int_0^{+\infty} x\,\mathrm{d}x &= \lim_{z\to 0+}\int_z^1 x\,\mathrm{d}x + \lim_{y\to+\infty}\int_1^y x\,\mathrm{d}x = \\ &= \lim_{z\to 0+}\left[\tfrac{1}{2}x^2\right]_z^1 + \lim_{y\to+\infty}\left[\tfrac{1}{2}x^2\right]_1^y = \frac{1}{2}+(+\infty) = +\infty;\\ \int_{-\infty}^{+\infty} x\,\mathrm{d}x &= \lim_{z\to-\infty}\int_z^0 x\,\mathrm{d}x + \lim_{y\to+\infty}\int_0^y x\,\mathrm{d}x = \\ &= \lim_{z\to-\infty}\left[\tfrac{1}{2}x^2\right]_z^0 + \lim_{y\to+\infty}\left[\tfrac{1}{2}x^2\right]_0^y = -\infty+(+\infty),\end{aligned}$$

tento součet limit však není definovaný, a tedy zobecněný Riemannův integrál $\int_{-\infty}^{+\infty} x\,\mathrm{d}x$ neexistuje a rovnosti v posledním výpočtu nemají smysl.

Následující lemma ukazuje, že pro omezené funkce na omezených intervalech pojmy Riemannova integrálu a zobecněného Riemannova integrálu splývají.

Lemma 41. Nechť $a, b \in \mathbb{R}$, $a < b$, a funkce f je omezená na intervalu $[a,b]$. Jestliže existuje Riemannův integrál funkce f na každém podintervalu $[c,d] \subset (a,b)$, pak existuje i Riemannův integrál funkce f na intervalu $[a,b]$.

Proof. Nechť $M > 0$ je konstanta splňující $|f(x)| < M$ pro každé $x \in [a,b]$. Zvolme libovolné $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, a dále body $c, d \in (a,b)$ tak, aby $c < d$ a $c - a < \frac{\varepsilon}{8M}$, $b - d < \frac{\varepsilon}{8M}$. Podle předpokladu existuje Riemannův integrál $\int_c^d f$. Podle Lemmatu 24(ii) existuje dělení D' intervalu $[c,d]$ takové, že $\overline{S}(f,D') - \underline{S}(f,D') < \frac{\varepsilon}{2}$. Nechť D je dělení intervalu $[a,b]$, které vznikne přidáním bodů a, b k dělení D'. Pak

$$\begin{aligned}\overline{S}(f,D) &= (c-a)\sup_{[a,c]} f + \overline{S}(f,D') + (b-d)\sup_{[d,b]} f \le \\ &\le M(c-a) + \overline{S}(f,D') + M(b-d) < \overline{S}(f,D') + \frac{\varepsilon}{4}.\end{aligned}$$

Obdobným způsobem obdržíme $\underline{S}(f,D) > \underline{S}(f,D') - \frac{\varepsilon}{4}$. Celkově tedy dostáváme $\overline{S}(f,D) - \underline{S}(f,D) < \overline{S}(f,D') - \underline{S}(f,D') + \frac{\varepsilon}{2} < \varepsilon$, což podle Lemmatu 24(ii) znamená, že existuje Riemannův integrál $\int_a^b f$. ∎

Může se stát, že funkce f má Riemannův integrál na všech podintervalech intervalu (a, b) a přesto nemá zobecněný Riemannův integrál (viz Příklad 40). Pro nezáporné funkce ovšem tato potíž nevzniká.

Lemma 42. Nechť $a, b \in \mathbb{R}^*$, $a < b$, f je nezáporná na (a, b) a f má Riemannův integrál na každém podintervalu $[x, y] \subset (a, b)$. Potom f má zobecněný Riemannův integrál na (a, b).

Proof. Zvolme $c \in (a, b)$. Je-li $c < x < y < b$, pak z Vět 25(ii) a 27(i) plyne

$$\int_c^y f = \int_c^x f + \int_x^y f \geq \int_c^x f,$$

což znamená, že funkce $y \mapsto \int_c^y f$ je neklesající na intervalu (c, b). Podle věty o limitě monotónní funkce (Věta **??**) tedy existuje $\lim_{y \to b-} \int_c^y f$. Podobně se lze přesvědčit o existenci $\lim_{x \to a+} \int_x^c f$. Podle Věty 27(i) a věty o limitě a uspořádání (Věta **??**) jsou obě limity nezáporné. Jejich součet je tedy definovaný, a tudíž existuje zobecněný Riemannův integrál f na (a, b). ∎

Pro zobecněný Riemannův integrál platí analogie některých vět o Riemannově integrálu.

Theorem 43. Nechť $a, b \in \mathbb{R}^*$ a $c \in (a, b)$.

(i) Jestliže funkce f má zobecněný Riemannův integrál na (a, b), pak má f zobecněný Riemannův integrál i na (a, c) a (c, b) a platí

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

(ii) Nechť funkce f má zobecněný Riemannův integrál na (a, c) a (c, b), f je omezená na nějakém okolí bodu c a součet $\int_a^c f + \int_c^b f$ má smysl. Pak f má zobecněný Riemannův integrál na (a, b) a platí

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Proof. (i) Zvolme $u \in (a, c)$. Z Lemmatu 39 plyne existence $\lim_{x \to a+} \int_x^u f$. Dále podle Lemmatu 38 platí $\lim_{y \to c-} \int_u^y f = \int_u^c f$, kde poslední integrál je Riemannův, a tedy reálné číslo. Zobecněný Riemannův integrál funkce f na (a, c) tedy existuje a platí

$$\int_a^c f = \lim_{x \to a+} \int_x^u f + \lim_{y \to c-} \int_u^y f = \lim_{x \to a+} \int_x^u f + \int_u^c f = \lim_{x \to a+} \int_x^c f,$$

přičemž poslední rovnost plyne z věty o aritmetice limit a z Věty 25(ii). Obdobně obdržíme

$$\int_c^b f = \lim_{y\to b-} \int_c^y f.$$

Tvrzení nyní snadno plyne z definice zobecněného Riemannova integrálu.

(ii) Nechť $u \in (a,c)$, $v \in (c,b)$ jsou takové body, že funkce f je omezená na $[u,v]$. Podle definice zobecněného Riemannova integrálu existují Riemannovy integrály funkce f na podintervalech $[x,y] \subset (a,c)$. Z Lemmatu 41 plyne, že existuje Riemannův integrál $\int_u^c f$ a Lemma 38 pak dává rovnost $\int_u^c f = \lim_{y\to c-} \int_u^y f$. Dostáváme tedy podobně jako v důkazu (i)

$$\int_a^c f = \lim_{x\to a+} \int_x^c f.$$

Analogicky ukážeme i

$$\int_c^b f = \lim_{y\to b-} \int_c^y f.$$

Právě provedené úvahy spolu s Větou 25 implikují existenci Riemannova integrálu funkce f na libovolném podintervalu $[x,y] \subset (a,b)$, odkud spolu s výše uvedenými rovnostmi plyne tvrzení věty. ∎

Theorem 44 (linearita zobecněného Riemannova integrálu)**.** Nechť $a,b \in \mathbb{R}^*$, $a<b$, f a g jsou funkce mající zobecněný Riemannův integrál na intervalu (a,b) a nechť $\alpha \in \mathbb{R}$. Potom

(i) funkce αf má zobecněný Riemannův integrál na (a,b) a platí

$$\int_a^b \alpha f = \alpha \int_a^b f,$$

má-li pravá strana smysl,

(ii) je-li součet $\int_a^b f + \int_a^b g$ definovaný, pak má funkce $f+g$ zobecněný Riemannův integrál na (a,b) a platí

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Tuto větu lze dokázat pomocí Lemmatu 39, Věty 26 a věty o aritmetice limit (Věta **??**).

Theorem 45. Nechť $a,b \in \mathbb{R}^*$, $a<b$, a nechť f a g jsou funkce mající zobecněný Riemannův integrál na intervalu (a,b). Potom platí:

(i) Je-li $f(x) \le g(x)$ pro každé $x \in (a,b)$, pak $\int_a^b f \le \int_a^b g$.

(ii) Funkce $|f|$ má zobecněný Riemannův integrál na intervalu (a,b) a platí $\left|\int_a^b f\right| \le \int_a^b |f|$.

Proof. (i) Zvolme pevně $c \in (a,b)$. Pro každé $y \in (c,b)$ platí dle Věty 27(i) $\int_c^y f \le \int_c^y g$. Použitím věty o limitě a uspořádání (Věta **??**) dostáváme nerovnost $\lim\limits_{y\to b-} \int_c^y f \le \lim\limits_{y\to b-} \int_c^y g$. Podobně ukážeme, že platí i nerovnost $\lim\limits_{x\to a+} \int_x^c f \le \lim\limits_{x\to a+} \int_x^c g$. Sečtením těchto nerovností dokážeme tvrzení (i).

(ii) Podle definice má funkce f Riemannův integrál na každém podintervalu $[x,y] \subset (a,b)$. Podle Věty 27(ii) má tedy také funkce $|f|$ Riemannův integrál na každém podintervalu $[x,y] \subset (a,b)$. Podle Lemmatu 42 tak existuje zobecněný Riemannův integrál funkce $|f|$ na intervalu (a,b). Zbytek tvrzení se dokáže analogicky jako v důkazu Věty 27(ii). ∎

Theorem 46. Nechť $a,b \in \mathbb{R}^*$, $a < b$, f je spojitá na (a,b), a F je primitivní funkce k f na (a,b). Pak zobecněný Riemannův integrál funkce f na (a,b) existuje, právě když existují limity $\lim\limits_{x\to a+} F(x)$ a $\lim\limits_{x\to b-} F(x)$ a jejich rozdíl má smysl. V tom případě platí

$$\int_a^b f = [F]_a^b = \lim_{x\to b-} F(x) - \lim_{x\to a+} F(x). \tag{18}$$

Proof. $\Rightarrow$ Zvolme $c \in (a,b)$. Pro libovolné $x \in (a,b)$ existuje Riemannův integrál $\int_c^x f$ a platí

$$\int_c^x f = [F]_c^x = F(x) - F(c), \tag{19}$$

kde první rovnost plyne z Newtonovy-Leibnizovy formule (Věta 31) a druhá ze spojitosti funkce F v bodech c a x. Použitím (19) dostaneme

$$\lim_{x\to b-} F(x) = F(c) + \lim_{x\to b-} \int_c^x f,$$
$$\lim_{x\to a+} F(x) = F(c) - \lim_{x\to a+} \int_x^c f.$$

Obě limity tedy existují, jejich rozdíl má smysl a platí vzorec (18).

$\Leftarrow$ Podobě jako v předchozí části důkazu vyjdeme ze vztahu (19) a limitním přechodem dokážeme požadovaná tvrzení. ∎

Pokud zobecněný Riemannův integrál funkce f na intervalu (a,b) existuje a přitom je konečný, pak říkáme, že $\int_a^b f$ **konverguje**. Pokud je roven $+\infty$ nebo $-\infty$, pak říkáme, že **diverguje**. Máme tedy následující možnosti:

$$\int_a^b f \begin{cases} \text{existuje a je roven} \begin{cases} \text{reálnému číslu, tj. konverguje,} \\ +\infty \text{ nebo } -\infty, \text{ tj. diverguje,} \end{cases} \\ \text{neexistuje.} \end{cases}$$

Example 47. Spočtěte integrál $\displaystyle\int_{-\infty}^{+\infty} \frac{1}{x^2+1}\,\mathrm{d}x$.

Solution. Integrovaná funkce je spojitá na $\mathbb{R}$, primitivní funkcí k ní je funkce arctg. Podle Věty 46 tedy dostáváme

$$\int_{-\infty}^{+\infty} \frac{1}{x^2+1}\,\mathrm{d}x = [\operatorname{arctg} x]_{-\infty}^{+\infty} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

♣

Example 48. Spočtěte integrály $\displaystyle\int_0^{+\infty} \frac{x}{x^2+1}\,\mathrm{d}x$ a $\displaystyle\int_{-\infty}^{+\infty} \frac{x}{x^2+1}\,\mathrm{d}x$.

Solution. Integrovaná funkce je spojitá na $\mathbb{R}$ a primitivní funkcí k ní je funkce $F(x) = \frac{1}{2}\log(x^2+1)$. Platí

$$F(0) = 0, \quad \lim_{x\to+\infty} F(x) = \lim_{x\to-\infty} F(x) = +\infty.$$

Z Věty 46 pak plyne, že $\int_0^{+\infty} \frac{x}{x^2+1}\,\mathrm{d}x = +\infty$ a $\int_{-\infty}^{+\infty} \frac{x}{x^2+1}\,\mathrm{d}x$ neexistuje. ♣

Example 49. Spočtěte $\displaystyle\int_0^1 \log x\,\mathrm{d}x$.

Solution. Funkce $\log$ je spojitá na intervalu $(0,+\infty)$. Primitivní funkci k ní spočteme pomocí metody per partes

$$\int \log x\,\mathrm{d}x = x\log x - \int x\cdot\frac{1}{x}\,\mathrm{d}x \overset{c}{=} x\log x - x.$$

Hodnota uvedené primitivní funkce v bodě 1 je -1 a dále podle Příkladu **??** platí

$$\lim_{x\to 0+}(x\log x - x) = 0.$$

Odtud a z Věty 46 dostáváme, že $\int_0^1 \log x\,\mathrm{d}x = -1$. ♣

Example 50. Integrál $\displaystyle\int_1^{+\infty} x^\alpha\,\mathrm{d}x$ konverguje, právě když $\alpha < -1$.

Proof. Pro $\alpha \neq -1$ je primitivní funkcí k funkci x^α na intervalu $(1,+\infty)$ funkce $\frac{x^{\alpha+1}}{\alpha+1}$. S pomocí Věty 46 dostaneme

$$\int_1^{+\infty} x^\alpha\,\mathrm{d}x = \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_1^{+\infty} = \begin{cases} 0 - \frac{1}{\alpha+1} = -\frac{1}{\alpha+1} & \text{pro } \alpha < -1, \\ +\infty - \frac{1}{\alpha+1} = +\infty & \text{pro } \alpha > -1. \end{cases}$$

Pro $\alpha = -1$ platí

$$\int_1^{+\infty} x^\alpha\,\mathrm{d}x = \int_1^{+\infty} \frac{1}{x}\,\mathrm{d}x = [\log x]_1^{+\infty} = +\infty - 0 = +\infty.$$

Je vhodné porovnat tento příklad s Větou **??**. ■

Example 51. Integrál $\int_0^1 x^\alpha \,\mathrm{d}x$ konverguje, právě když $\alpha > -1$.

Proof. Pro $\alpha \neq -1$ je primitivní funkcí k funkci x^α na intervalu $(0,1)$ funkce $\frac{x^{\alpha+1}}{\alpha+1}$. S pomocí Věty 46 dostaneme

$$\int_0^1 x^\alpha \,\mathrm{d}x = \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_0^1 = \begin{cases} \frac{1}{\alpha+1} - 0 = \frac{1}{\alpha+1} & \text{pro } \alpha > -1, \\ \frac{1}{\alpha+1} - (-\infty) = +\infty & \text{pro } \alpha < -1. \end{cases}$$

Pro $\alpha = -1$ platí

$$\int_0^1 x^\alpha \,\mathrm{d}x = \int_0^1 \frac{1}{x} \,\mathrm{d}x = [\log x]_0^1 = 0 - (-\infty) = +\infty. \qquad \blacksquare$$

Remark. Podobně jako v předchozím příkladu lze ukázat, že je-li $c, d \in \mathbb{R}$, $c < d$, pak $\int_c^d (x-c)^\alpha \,\mathrm{d}x$ konverguje, právě když platí $\alpha > -1$. Stejně tak pro $d < c$ integrál $\int_d^c (c-x)^\alpha \,\mathrm{d}x$ konverguje, právě když $\alpha > -1$. Toto pozorování využijeme později v konkrétních příkladech.

U řady integrálů poznáme, zda konvergují, pokud je porovnáme s vhodnou funkcí $x \mapsto x^\alpha$. K tomu nám poslouží následující dvě věty.

Theorem 52 (srovnávací kritérium)**.** Nechť $a, b \in \mathbb{R}^*$, $a < b$, funkce f a g splňují $0 \leq f(x) \leq g(x)$ pro všechna $x \in (a,b)$ a f je na (a,b) spojitá. Pokud konverguje $\int_a^b g$, pak konverguje i $\int_a^b f$.

Proof. Ze spojitosti funkce f a z Lemmatu 42 plyne existence zobecněného Riemannova integrálu $\int_a^b f$. Podle Věty 45(i) potom tento integrál konverguje. $\blacksquare$

Theorem 53 (limitní srovnávací kritérium)**.** Nechť f a g jsou spojité nezáporné funkce na intervalu $[a,b)$, $b \in \mathbb{R}^*$, a existuje limita $\lim_{x \to b-} \frac{f(x)}{g(x)} = \gamma \in \mathbb{R}^*$.

- Je-li $\gamma \in (0, +\infty)$, pak $\int_a^b f$ konverguje, právě když konverguje $\int_a^b g$.
- Je-li $\gamma = 0$, pak z konvergence $\int_a^b g$ plyne konvergence $\int_a^b f$.
- Je-li $\gamma = +\infty$, pak z divergence $\int_a^b g$ plyne divergence $\int_a^b f$.

Proof. Předpokládejme nejprve, že $\gamma \in \langle 0, +\infty)$. Z definice limity plyne, že existuje takové $c \in \mathbb{R}$, že pro všechna $x \in (c,b)$ je $\left|\frac{f(x)}{g(x)} - \gamma\right| < 1$. Speciálně pro $x \in (c,b)$ máme $0 \leq \frac{f(x)}{g(x)} < \gamma + 1$, neboli $0 \leq (x) < (\gamma+1)g(x)$. Předpokládáme-li, že $\int_a^b g$ konverguje, pak z Věty 43(i) plyne konvergence integrálu $\int_c^b g$. Konverguje tedy také $\int_c^b (\gamma+1)g$ (Věta 44(i)), a proto dle Věty 52 konverguje i integrál $\int_c^b f$. Funkce f je spojitá na $[a,c]$, a tedy podle Věty 29 integrál $\int_a^c f$ konverguje. Podle Věty 43(ii) tak konverguje i integrál $\int_a^b f$. Odtud plyne druhý bod tvrzení a jedna implikace v prvním bodě.

Nyní předpokládejme, že platí $\gamma \in (0, +\infty) \cup \{+\infty\}$. Potom máme $\lim_{x \to b-} \frac{g(x)}{f(x)} \in [0, +\infty)$, čili podle již dokázaného platí, že konverguje-li $\int_a^b f$, pak konverguje i $\int_a^b g$. Odtud plyne třetí bod tvrzení a zbývající implikace v prvním bodě. ■

Remarks. 1. Analogická tvrzení platí pro funkce na intervalu $(a, b]$.

2. Porovnejte srovnávací kritérium a jeho limitní verzi s analogickými kritérii pro konvergenci řad uvedenými v kapitole **??**.

Ukažme si několik příkladů, jak tyto věty používat.

Example 54. Zjistěte, zda $\displaystyle\int_1^{+\infty} x^{20} e^{-x^2} \, \mathrm{d}x$ konverguje.

Solution. Funkce $f(x) = x^{20} e^{-x^2}$ je spojitá a kladná na intervalu $[1, +\infty)$. Víme, že $\int_1^{+\infty} \frac{1}{x^2} \, \mathrm{d}x$ konverguje (Příklad 50). Dále platí

$$\lim_{x \to +\infty} \frac{f(x)}{\frac{1}{x^2}} = \lim_{x \to +\infty} \frac{x^{22}}{e^{x^2}} = 0,$$

a tedy podle Věty 53 integrál ze zadání konverguje. ♣

Example 55. Zjistěte, zda $\displaystyle\int_1^{+\infty} \frac{x^{17} + 36}{x^{18} + 51x^7 + 5} \, \mathrm{d}x$ konverguje.

Solution. Integrovaná funkce je spojitá a kladná na intervalu $[1, +\infty)$. Protože stupeň čitatele je o jedna menší než stupeň jmenovatele, je vhodné srovnat integrovanou funkci s funkcí $1/x$:

$$\lim_{x \to +\infty} \frac{\frac{x^{17}+36}{x^{18}+51x^7+5}}{\frac{1}{x}} = \lim_{x \to +\infty} \frac{x^{18} + 36x}{x^{18} + 51x^7 + 5} = 1.$$

Podle Věty 53 integrál ze zadání konverguje, právě když konverguje integrál $\int_1^{+\infty} \frac{1}{x} \, \mathrm{d}x$. Tento integrál však diverguje podle Příkladu 50. Tudíž i zadaný integrál diverguje. ♣

Example 56. Zjistěte, pro které hodnoty $\alpha \in \mathbb{R}$ konverguje $\displaystyle\int_0^{\pi} \sin^{\alpha} x \, \mathrm{d}x$.

Solution. Integrovaná funkce je spojitá a kladná na intervalu $(0, \pi)$. Rozdělme interval $(0, \pi)$ na dvě části – zkoumejme $\int_0^{\pi/2} \sin^{\alpha} x \, \mathrm{d}x$ a $\int_{\pi/2}^{\pi} \sin^{\alpha} x \, \mathrm{d}x$, neboť integrál ze zadání konverguje, právě když konvergují oba uvedené integrály přes menší intervaly (Věta 43).

Funkce $x \mapsto \sin^{\alpha} x$ je spojitá na intervalu $(0, \pi/2]$ a $\lim_{x \to 0+} \frac{\sin^{\alpha} x}{x^{\alpha}} = 1$. Tedy $\int_0^{\pi/2} \sin^{\alpha} x \, \mathrm{d}x$ konverguje, právě když konverguje $\int_0^{\pi/2} x^{\alpha} \, \mathrm{d}x$. Tento integrál konverguje, právě když $\alpha > -1$ (Příklad 51 a poznámka za ním).

Zbývá vyšetřit $\int_{\pi/2}^{\pi} \sin^\alpha x \,\mathrm{d}x$. Protože však $\sin(\pi - x) = \sin x$, je tento integrál roven integrálu z předchozího odstavce, a tedy konverguje, právě když $\alpha > -1$. Integrál ze zadání tedy konverguje, právě když $\alpha > -1$. ♣

3.4. Cvičení

K zadané funkci nalezněte na co největších intervalech nějakou primitivní funkci F.

1. $(x^2 - x)\exp x$

2. $5^x \sin x$

3. $\dfrac{\log^2 x}{x}$

4. $\dfrac{x}{\sqrt{4 - x^4}}$

5. $\dfrac{1}{3x^2 - 2x - 1}$

6. $\dfrac{2x}{(x+1)(x^4 + 2x^2 + 1)}$

7. $\dfrac{\operatorname{cotg} x}{\sin x + \cos x - 1}$

8. $\dfrac{\sin x}{\sin^3 x + \cos^3 x}$

9. $\dfrac{\sqrt{2x+1}}{x^2}$

10. $\dfrac{x+2}{(x^2 + x + 1)^2 (x-1)}$

11. $\dfrac{1}{(x+1)\sqrt{x^2 + x + 1}}$

12. $\dfrac{1}{x\sqrt{2 + x - x^2}}$

Spočtěte následující určité integrály.

13. $\displaystyle\int_0^{\pi/4} \frac{\sin x - \cos x}{\sin x - 2\cos^3 x}\,\mathrm{d}x$

14. $\displaystyle\int_{-1/2}^{1} \frac{1}{\sqrt{8 + 2x - x^2}}\,\mathrm{d}x$

15. $\displaystyle\int_0^{\pi} x^2 \sin^2 x \,\mathrm{d}x$

16. $\displaystyle\int_0^{1/2} \arccos x \,\mathrm{d}x$

17. $\displaystyle\int_{-1}^{1} \frac{x}{\sqrt{5 - 4x}}\,\mathrm{d}x$

18. $\displaystyle\int_0^{1} x \log^2 x \,\mathrm{d}x$

19. $\displaystyle\int_0^{+\infty} x^2 e^{-x}\,\mathrm{d}x$

20. $\displaystyle\int_{-\infty}^{+\infty} x^2 e^{-x}\,\mathrm{d}x$

21. $\displaystyle\int_{-\infty}^{+\infty} x^3 e^{-x^2}\,\mathrm{d}x$

22. $\displaystyle\int_0^{+\infty} x^3 e^{-x^2}\,\mathrm{d}x$

23. Vypočítejte obsah obrazce ohraničeného grafy dvou funkcí $x \mapsto \frac{2}{1+x^2}$ a $x \mapsto x^2$.

24. Vypočítejte obsah obrazce ohraničeného grafy funkcí $x \mapsto x^2 - 6x + 8$, $x \mapsto -4x + 7$ a $x \mapsto 2x - 8$.

25. Vypočítejte délku křivky, která je grafem funkce $f(x) = \log(\cos x)$ pro $x \in [0, \pi/6]$.

26. Vypočítejte obsah rotační plochy, která vznikne rotací křivky $x \mapsto x^2/2$, $x \in [0, 3/4]$, kolem osy x.

27. Vypočítejte objem rotačního tělesa, které vznikne rotací obrazce ležícího v rovině x, y kolem osy x. Obrazec je ohraničen křivkami, jejichž rovnice jsou $x^2 - \frac{1}{2}y^2 = 1$ a $y^2 - x^2 = 1$.

28. Pro které hodnoty parametrů $p, q \in \mathbb{R}$ konverguje $\displaystyle\int_0^1 x^p(1-x)^q \,\mathrm{d}x$?

29. Zjistěte, pro které hodnoty parametrů $\alpha, \beta, \gamma \in \mathbb{R}$ konverguje integrál

$$\int_0^{\frac{\pi}{2}} \sin^\alpha x \cos^\beta x (1 - \cos x)^\gamma \,\mathrm{d}x.$$

Výsledky cvičení

1. $F(x) = (x^2-3x+3)\exp x$ na celém $\mathbb{R}$ **2.** $F(x) = \frac{5^x}{1+\log^2 5}(\log 5 \cdot \sin x - \cos x)$ na celém $\mathbb{R}$ **3.** $F(x) = \frac{1}{3}\log^3 x$ na intervalu $(0, +\infty)$ **4.** $F(x) = \frac{1}{2}\arcsin(\frac{1}{2}x^2)$ na intervalu $(-\sqrt{2}, \sqrt{2})$ **5.** $F(x) = \frac{1}{4}\log\left|\frac{x-1}{3x+1}\right|$ na každém z intervalů $(-\infty, -1/3)$, $(-1/3, 1)$ a $(1, +\infty)$ **6.** $F(x) = -\frac{1}{2}\log|x+1| + \frac{1}{4}\log(x^2+1) + \frac{1}{2}\frac{x-1}{x^2+1}$ na intervalech $(-\infty, -1)$ a $(-1, +\infty)$ **7.** $F(x) = -\frac{1}{2}\operatorname{cotg}\frac{x}{2} + \frac{1}{2}\log\left|\operatorname{tg}\frac{x}{2}\right|$ na libovolném intervalu neobsahujícím body množiny $\{k\pi;\ k \in \mathbb{Z}\} \cup \{\pi/2 + 2k\pi;\ k \in \mathbb{Z}\}$ **8.** $F(x) = -\frac{1}{3}\log|\operatorname{tg} x + 1| + \frac{1}{6}\log(\operatorname{tg}^2 x - \operatorname{tg} x + 1) + \frac{1}{\sqrt{3}}\operatorname{arctg}\frac{2\operatorname{tg} x - 1}{\sqrt{3}}$ na libovolném intervalu neobsahujícím body množiny $\{(2k+1)\pi/2;\ k \in \mathbb{Z}\} \cup \{-\pi/4 + k\pi;\ k \in \mathbb{Z}\}$ **9.** $F(x) = \log\frac{\left|\sqrt{2x+1}-1\right|}{\sqrt{2x+1}+1} - \frac{\sqrt{2x+1}}{x}$ na intervalech $(-1/2, 0)$ a $(0, +\infty)$ **10.** $F(x) = \frac{1}{3}\log|x-1| - \frac{1}{6}\log(x^2+x+1) - \frac{5\sqrt{3}}{9}\operatorname{arctg}\left(\frac{2x+1}{\sqrt{3}}\right) - \frac{1}{3}\cdot\frac{x-1}{x^2+x+1}$ na intervalech $(-\infty, 1)$ a $(1, +\infty)$ **11.** $F(x) = \log\left|\frac{\sqrt{x^2+x+1}-x-2}{\sqrt{x^2+x+1}-x}\right|$ na intervalech $(-\infty, -1)$ a $(-1, +\infty)$ **12.** $F(x) = \frac{1}{\sqrt{2}}\log\left|\frac{\sqrt{\frac{x+1}{2-x}}-\frac{1}{\sqrt{2}}}{\sqrt{\frac{x+1}{2-x}}+\frac{1}{\sqrt{2}}}\right|$ na intervalech $(-1, 0)$ a $(0, 2)$ **13.** $\frac{2}{\sqrt{7}}\left(\operatorname{arctg}\frac{3}{\sqrt{7}} - \operatorname{arctg}\frac{1}{\sqrt{7}}\right)$ **14.** $\pi/6$ **15.** $\frac{\pi^3}{6} - \frac{\pi}{4}$ **16.** $\frac{\pi}{6} - \frac{\sqrt{3}}{2} + 1$

17. $1/6$ **18.** $1/4$ **19.** 2 **20.** $+\infty$ **21.** 0 **22.** $1/2$ **23.** $\pi - 2/3$ **24.** $9/4$ **25.** $\frac{1}{2}\log 3$ **26.** $\frac{\pi}{16}\left(\frac{255}{64} - 2\log 2\right)$ **27.** $\frac{4}{3}\pi\left(3\sqrt{3} - 2\right)$ **28.** Integrál konverguje, právě když $p > -1$ a $q > -1$. **29.** Integrál konverguje, právě když $\beta > -1$ a $\alpha + 2\gamma > -1$.

V. Hájková, M. Johanis, O. John, O. Kalenda, M. Zelený

MATHEMATICS

Translation D. Campbell

Vydal
MATFYZPRESS
vydavatelství Matematicko-fyzikální fakulty
Univerzity Karlovy v Praze,
Sokolovská 83, 186 75 Praha 8
jako svou 387. publikaci.

Obrázky byly vytvořeny v programech
Maple™ a CorelDRAW®.
Na obálce je použita reprodukce mědirytiny
Albrechta Dürera Melancholie I.

Druhé upravené vydání

Praha 2012

ISBN 978-80-7378-193-7
ISBN 80-86732-99-1 (1. vydání)