Lecture Notes for Mathematics for Physicists 1 & 2

Dr. Stephan Mescher

University of Leipzig, winter term 2020/2021 & summer term 2021
(current version: July 6, 2026)

Contents

Chapter 1

Sets and numbers

1.1 Basic set theory

In this course, we aim for a rigorous formal description of all mathematical notions that we will study. However, as a very first step, we need to let one notion fall from the sky, namely the notion of a *set*.

A set *is a well-defined collection of distinct objects of our perception or of our thought. These objects are called* elements *of the set.*[1]

Well-defined means: for each object it can be uniquely determined whether it belongs to the set or not.
Distinct means: every element occurs only once in the set.

We write $x \in M$ if x is an element of the set M and $x \notin M$ if x is not an element of M. A set can be written as $\{\dots\}$, where the dots are replaced by a listing or a description of its elements.

Example 1.1. (1) $\{1, 2, 3\}$ is a set with three elements.

(2) $\mathbb{N} = \{1, 2, 3, 4, \dots\}$, the set of *natural numbers*.[2]

(3) $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \dots\}$, the set of *integers*.

(4) $2\mathbb{Z} = \{0, 2, -2, 4, -4, 6, -6, \dots\}$, the set of *even integers*.
The notation ":=" means that the object on the left of the symbol *is defined as* the object on the right of the symbol.

All sets in the previous example are defined using a roster notation (or enumeration notation), i.e. by simply listing their elements.
Sets can also be described using *propositional notation*. If $A(x)$ is a logical proposition making a statement about some object x that can be uniquely verified as true or false, then we let

$$\{x \mid A(x)\}$$

[1] This description goes back to Georg Cantor (1845-1918), one of the founders of set theory who was a professor at the University of Halle (which is only 35 km from Leipzig).

[2] Please be aware that there are two different conventions in use: in this lecture it holds that $\mathbb{N} = \{1, 2, 3, \dots\}$, while other books or lecture notes consider $\{0, 1, 2, 3, \dots\}$ as the set of natural numbers.

denote the set of all objects x for which $A(x)$ is true. If we consider only objects that lie in a given set M we can also write

$$\{x \in M \mid A(x)\}$$

to determine a set.

Example 1.2. (1) $M_1 = \{x \in \mathbb{Z} \mid x > 5\} = \{6, 7, 8, \dots\}$.

(2) $M_2 = \{x \in \mathbb{Z} \mid \frac{1}{2}x \in \mathbb{Z}\} = 2\mathbb{Z}$.

(3) $M_3 = \{x \in \mathbb{Z} \mid 3x = 2\} = \varnothing$. Here, $\varnothing$ denotes the set which does not contain any element. This set is called the *empty set*.

Definition 1.3. Let M and N be sets.

a) $M \cup N := \{x \mid x \in M \text{ or } x \in N\}$ is the *union of M and N*.

b) $M \cap N := \{x \mid x \in M \text{ and } x \in N\}$ is the *intersection of M and N*. M and N are called *disjoint* if $M \cap N = \varnothing$.

c) $M \setminus N := \{x \mid x \in M \text{ and } x \notin N\}$ is the *difference of M and N*.

Example 1.4. If

$$M = \{1, \text{tree}, 17, \text{Leipzig}\}, \quad N = \{1, 2, \text{tree}, 27, \text{Dresden}\},$$

then

$$\begin{aligned} M \cup N &= \{1, 2, \text{tree}, 17, 27, \text{Leipzig}, \text{Dresden}\}, \\ M \cap N &= \{1, \text{tree}\}, \\ M \setminus N &= \{17, \text{Leipzig}\}. \end{aligned}$$

We can generalize the notions of intersection and union to arbitrarily many sets instead of just two. Before we do so, we introduce some notation that we will make heavy use of in the following chapters.

Definition 1.5. Let M be a set and $A(x)$ be a logical proposition making a statement about $x \in M$.

a) We write

$$\forall x \in M: \ \ A(x) \qquad \text{or} \qquad A(x) \ \forall x \in M$$

to express that $A(x)$ is fulfilled *for all* $x \in M$.

b) We write

$$\exists x \in M: \ \ A(x)$$

to express that *there exists* an $x \in M$ for which $A(x)$ is true.

c) If A and B are logical propositions, then we write

$$A \Rightarrow B$$

and say that A *implies* B if whenever A is true it follows that B is true. (This does not say anything about A being true or false, it just means that *if* A is true, *then* B will be true as well.

d) We say that the logical propositions A and B are *equivalent* and write

$$A \Leftrightarrow B$$

if $A \Rightarrow B$ and $B \Rightarrow A$, i.e. in the case that A is true *if and only if* B is true. (Again, this does not make any claim about the statements being true or false, it is just a relation between both statements).

e) We write $\wedge$ for the term "and" between two logical statements.

f) We write $\vee$ for the term "or" between two logical statements.

Example 1.6.

$$\text{Anna lives in Leipzig} \quad \Rightarrow \quad \text{Anna lives in Germany}$$
$$(x \in M \vee x \in N) \quad \Leftrightarrow \quad x \in M \cup N$$

Definition 1.7. Let I be a set and let $(M_i)_{i\in I}$ be a family of sets, i.e. for each $i \in I$ there is a given set M_i. The *union of* $(M_i)_{i\in I}$ is given by

$$\bigcup_{i\in I} M_i := \{x \mid \exists i \in I : \ x \in M_i\}.$$

The *intersection of* $(M_i)_{i\in I}$ is given by

$$\bigcap_{i\in I} M_i := \{x \mid \forall i \in I : \ x \in M_i\}.$$

Example 1.8. Let $I = \mathbb{N}$ and let $A_n := \{1, 2, \ldots, n\}$ for each $n \in \mathbb{N}$. Then

$$\bigcup_{n\in\mathbb{N}} A_n = \mathbb{N} \qquad \text{and} \qquad \bigcap_{n\in\mathbb{N}} A_n = \{1\}.$$

We continue by formalizing the notion of a subset of a set and study some related terminology.

Definition 1.9. Let M and N be sets.

a) M is called a *subset of* N, denoted by $M \subset N$, if every element of M is also an element of N.[3] We also write $N \supset M$ if and only if $M \subset N$.

b) M and N are called *equal*, denoted by $M = N$, if $M \subset N$ and $N \subset M$.

c) If $M \subset N$, then we denote $M^c := N \setminus M$ and call it *the complement of M in N*.

d) The set

$$\mathcal{P}(M) = \{A \mid A \subset M\},$$

i.e. the set of all subsets of M, is called *the power set of M*.

Remark 1.10. If one wants to prove that two sets M and N are equal, but can not find a direct reformulation of M as N, it is often recommended to follow the definition and prove the two statements $M \subset N$ and $M \supset N$ individually. We will see examples for this method later on.

[3] In many books, the symbol $\subseteq$ is used instead of $\subset$ to indicate a subset.

Example 1.11. If $M = \{1,2,3\}$ then

$$\mathcal{P}(M) = \{\{1,2,3\},\{1,2\},\{1,3\},\{2,3\},\{1\},\{2\},\{3\},\varnothing\}.$$

For $A = \{1,2\}$ it holds that $A \subset M$, but $A \neq M$. Moreover, $A^c = \{3\}$.

The following theorem collects several basic facts about unions, intersections and complements. We will only prove a small part of it and the reader is encouraged to try to work out proofs of the remaining parts in a similar way.

Theorem 1.12. *Let M_1, M_2 and M_3 be sets. Then:*

a) (Commutativity) $M_1 \cup M_2 = M_2 \cup M_1$ *and* $M_1 \cap M_2 = M_2 \cap M_1$.

b) (Associativity)

$$M_1 \cup (M_2 \cup M_3) = (M_1 \cup M_2) \cup M_3 \quad \textit{and} \quad M_1 \cap (M_2 \cap M_3) = (M_1 \cap M_2) \cap M_3.$$

c) (Distributivity)

$$\begin{aligned} &(1) \quad M_1 \cup (M_2 \cap M_3) = (M_1 \cup M_2) \cap (M_1 \cup M_3), \\ &(2) \quad M_1 \cap (M_2 \cup M_3) = (M_1 \cap M_2) \cup (M_1 \cap M_3). \end{aligned}$$

d) (De Morgan's laws) *If $M_1 \subset M_3$ and $M_2 \subset M_3$, then*

$$M_1^c \cup M_2^c = (M_1 \cap M_2)^c \qquad \textit{and} \qquad M_1^c \cap M_2^c = (M_1 \cup M_2)^c.$$

Proof of part c).(1). We will follow the advice from Remark 1.10 and prove the two inclusions individually.

"$\subset$": We want to show that $M_1 \cup (M_2 \cap M_3) \overset{!}{\subset} (M_1 \cup M_2) \cap (M_1 \cup M_3)$.
If $x \in M_1 \cup (M_2 \cap M_3)$, then $x \in M_1$ or $x \in M_2 \cap M_3$. In the case that $x \in M_1$, we obtain:

$$x \in M_1 \;\Rightarrow\; x \in M_1 \cup M_2 \wedge x \in M_1 \cup M_3 \;\Leftrightarrow x \in (M_1 \cup M_2) \cap (M_1 \cup M_3)$$

by definition of the intersection.
In the case that $x \in M_2 \cap M_3$, we derive:

$$x \in M_2 \wedge x \in M_3 \;\Rightarrow x \in M_1 \cup M_2 \wedge x \in M_1 \cup M_3 \;\Leftrightarrow x \in (M_1 \cup M_2) \cap (M_1 \cup M_3).$$

Combining both cases, we have shown the claim.

"$\supset$": We want to show that $(M_1 \cup M_2) \cap (M_1 \cup M_3) \overset{!}{\subset} M_1 \cup (M_2 \cap M_3)$.
If $x \in (M_1 \cup M_2) \cap (M_1 \cup M_3)$, then $x \in M_1 \cup M_2$ and $x \in M_1 \cup M_3$. In the case that $x \in M_1$, it follows that $x \in M_1 \cup (M_2 \cap M_3)$. If $x \notin M_1$, it follows that $x \in M_2$ and $x \in M_3$, hence $x \in M_2 \cap M_3$, which implies $x \in M_1 \cup (M_2 \cap M_3)$. Combining both cases shows the claim. □

Definition 1.13. Let M and N be sets. The *Cartesian product of M and N* is the set

$$M \times N = \{(x,y) \mid x \in M \text{ and } y \in N\}.$$

More generally, if $M_1, M_2, \ldots, M_n$ are sets, where $n \in \mathbb{N}$, the Cartesian product of $M_1, M_2, \ldots, M_n$ is defined as

$$\prod_{i=1}^{n} M_i := M_1 \times M_2 \times \cdots \times M_n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in M_i \; \forall i \in \{1,2,\ldots,n\}\}.$$

Example 1.14. (1) If $M = \{1,2\}$, $N = \{0,1\}$, then $M \times N = \{(1,0),(2,0),(1,1),(2,1)\}$.

(2) Assume that the real numbers $\mathbb{R}$ have already been constructed. Then

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x,y) \mid x,y \in \mathbb{R}\}$$

is geometrically interpreted as the Euclidean plane, while

$$\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R} = \{(x,y,z) \mid x,y,z \in \mathbb{R}\}.$$

is seen as three-dimensional Euclidean space, e.g. in classical mechanics.

Definition 1.15. Let X and Y be sets.

a) A *map* $f : X \to Y$ associates with every element of X a unique element of Y. The element associated with $x \in X$ by f is denoted by $f(x) \in Y$. If Y is a set of numbers, f is also called a *function*.

b) X is called the *domain of* f and Y is called the *codomain* or *set of destination of* f.

c) Two maps $f, g : X \to Y$ are called *equal* if

$$f(x) = g(x) \qquad \forall x \in X.$$

d) For $A \subset X$ the set $f(A) := \{y \in Y \mid \exists x \in A : \ f(x) = y\}$ is called *the image of* A *under* f. For $B \subset Y$ the set $f^{-1}(B) := \{x \in X \mid f(x) \in B\}$ is called *the preimage of* B *under* f.

Example 1.16. The following are examples of maps:

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2,$$
$$g : \mathbb{R}^2 \to \mathbb{R}, \quad g(x,y) = \sqrt{x^2 + y^2},$$
$$h : \{\text{students at the University of Leipzig}\} \to \mathbb{R}, \qquad h(x) := \text{the height of person } x.$$

The following theorem summarizes the behaviour of images and preimages under taking unions intersections of sets. Again, we will only prove one part of it and encourage the reader to try to prove the rest on their own.

Theorem 1.17. *Let X and Y be sets and $f : X \to Y$ be a map.*

a) For all $A_1, A_2 \subset X$ it holds that

(i) $f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$,
(ii) $f(A_1 \cap A_2) \subset f(A_1) \cap f(A_2)$,
(iii) $A_1 \subset f^{-1}(f(A_1))$.

b) For all $B_1, B_2 \subset Y$ it holds that

(i) $f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$,
(ii) $f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2)$.

Proof of part a).(i). Here, we can prove the equation directly by reformulations the set according to the definition.

$$\begin{aligned} f(A_1 \cup A_2) &= \{y \in Y \mid \exists x \in A_1 \cup A_2 : \ f(x) = y\} \\ &= \{y \in Y \mid (\exists x \in A_1 : \ f(x) = y) \ \vee \ (\exists x \in A_2 : \ f(x) = y)\} \\ &= \{y \in Y \mid \exists x \in A_1 : \ f(x) = y\} \cup \{y \in Y \mid \exists x \in A_2 : \ f(x) = y\} \\ &= f(A_1) \cup f(A_2). \end{aligned}$$

□

Definition 1.18. Let X, Y and Z be sets and $f : X \to Y$ and $g : Y \to Z$ be maps. The map

$$g \circ f : X \to Z, \qquad (g \circ f)(x) := g(f(x)),$$

is called *the composition of g and f.*

Example 1.19. Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 + x$, and $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$. Then $g \circ f : \mathbb{R} \to \mathbb{R}$ is given by

$$(g \circ f)(x) = g(f(x)) = (f(x))^2 = (x^2 + x)^2 = x^4 + 2x^3 + x^2.$$

The notions in the following definition are essential terms for the study of maps and will occur at various points throughout this course.

Definition 1.20. Let X and Y be sets and let $f : X \to Y$ be a map.

a) f is *injective* if for all $x, y \in X$ with $x \neq y$ it holds that $f(x) \neq f(y)$, i.e. if no two distinct elements of X are mapped to the same element of Y.

b) f is *surjective* if for each $y \in Y$ there exists an $x \in X$ with $f(x) = y$, i.e. if $f(X) = Y$.

c) f is called *bijective* if it is injective and surjective, i.e. if for each $y \in Y$ there exists a *unique* $x \in X$ with $f(x) = y$.

Example 1.21. (1) $f : \mathbb{N} \to \mathbb{N}$, $f(n) = n + 1$, is injective since $f(m) = f(n)$ is fulfilled if and only if $m = n$. However, f is not surjective since there is no $n \in \mathbb{N}$ with $f(n) = 1$.

(2) $g : \mathbb{N} \to \mathbb{N}$, given by $g(n) = n - 1$ if $n > 1$ and $g(1) = 1$ is surjective, since for each $n \in \mathbb{N}$ it holds that $g(n + 1) = n$, hence $g(\mathbb{N}) = \mathbb{N}$. However, g is not injective since $f(1) = f(2) = 1$, but $1 \neq 2$.

(3) $h : \{1, 2, 3\} \to \{a, b, c\}$, given by $h(1) = a$, $h(2) = b$ and $h(3) = c$, is bijective.

If a map $f : X \to Y$ is bijective, it puts the elements of X and those of Y in a one-to-one relationship. Thus, instead of associating an element of Y with each element of X, we might just as well do it the other way around. This is formalized in the following theorem.

Theorem 1.22. *Let X and Y be sets and let $f : X \to Y$ be a bijective map. Then there exists a unique map $g : Y \to X$, such that*

$$(g \circ f)(x) = x \ \ \forall x \in X \qquad \textit{and} \qquad (f \circ g)(y) = y \ \ \forall y \in Y.$$

Proof. Uniqueness: Assume $g_1, g_2 : Y \to X$ are maps with $(g_1 \circ f)(x) = (g_2 \circ f)(x)$ for all $x \in X$. Let $y \in Y$ be arbitrary. Since f is surjective, there exists an $x \in X$ with $f(x) = y$. Consequently,

$$g_1(y) = g_1(f(x)) = g_2(f(x)) = g_2(y) \quad \forall y \in Y,$$

which means $g_1 = g_2$.
Existence: Let $y \in Y$ be arbitrary. Since f is bijective, there exists a unique x with $f(x) = y$, so we can put $g(y) := x$. One checks that $(g \circ f)(x) = x$ and $(f \circ g)(y) = y$. Defining $g(y)$ for each $y \in Y$ in this way, we obtain a well-defined map with the desired properties. □

Remark 1.23. In the situation of Theorem 1.22 we put $f^{-1} := g$ and call it the *inverse map of* f.

1.2 Groups and fields

Our next main aim is the construction of the real numbers which we have already encountered in high school. On the real numbers we have two basic operations on which all constructions involving them are based: addition and multiplication. (Subtraction is nothing but addition with negative numbers, while division is nothing but multiplication with reciprocals.)
In this section, we want to formalize these operations by introducing the notion of a *field*, which is roughly described as a set with an addition and a multiplication, which behave "reasonably" as we would intuitively expect them to behave from our experience with real numbers.

Before we study fields, we discuss the notion of a *group* which can be seen as the most elementary notion of a set which possesses a "reasonable" algebraic operation. Groups appear in physics whenever one studies physical systems *with symmetries*, e.g. a spinning top which is symmetrically shaped with respect to rotations.

Definition 1.24. A *group* is a pair $(G, \cdot)$, where G is a set and $\cdot : G \times G \to G$ is a map with the following properties:

(G1) *(Associativity)* For all $g_1, g_2, g_3 \in G$ it holds that

$$(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3).$$

(G2) *(Existence of a neutral element)* There exists an element $e \in G$, such that

$$g \cdot e = e \cdot g = g \quad \forall g \in G.$$

e is called a *neutral element* or *identity element of* G.

(G3) *(Existence of inverse elements)* For every $g \in G$ there exists an element $g^{-1} \in G$, such that

$$g \cdot g^{-1} = g^{-1} \cdot g = e.$$

g^{-1} is called *the inverse element of* g.

A group $(G, \cdot)$ is called *abelian* or *commutative* if

$$g \cdot h = h \cdot g \qquad \forall g, h \in G.$$

The map $\cdot$ is called the *group operation*. If it is obvious which group operation we are referring to, then we simply write G instead of $(G, \cdot)$.

Example 1.25. $(\mathbb{Z}, +)$ is an abelian group, where $+$ is the usual addition. The neutral element is given $e = 0$ while the inverse element of $n \in \mathbb{Z}$ is given by $-n$.

Remark 1.26. We can derive some properties of a group $(G, \cdot)$ directly from its defining axioms:

(1) The neutral element of G is *unique*: If $e, e' \in G$ are both neutral elements, then

$$e \overset{\text{(i)}}{=} e \cdot e' \overset{\text{(ii)}}{=} e',$$

where (i) follows from the neutrality of e' and (ii) follows from the neutrality of e.

(2) For each $g \in G$ there exists a *unique* inverse element: If $h, h' \in G$ are inverse elements of g, then

$$h' \overset{\text{(G2)}}{=} e \cdot h' \overset{(*)}{=} (h \cdot g) \cdot h' \overset{\text{(G1)}}{=} h \cdot (g \cdot h') \overset{(**)}{=} h \cdot e \overset{\text{(G2)}}{=} h,$$

where $(*)$ holds since h is an inverse element of g and $(**)$ holds since h' is an inverse element of g.

Theorem 1.27. *Let $(G, \cdot)$ be a group.*

a) For each $x \in G$ it holds that $(x^{-1})^{-1} = x$.

b) For all $x, y \in G$ it holds that $(x \cdot y)^{-1} = y^{-1} \cdot x^{-1}$.

Proof. a) By definition of inverse elements, it holds that $x^{-1} \cdot (x^{-1})^{-1} = (x^{-1})^{-1} \cdot x^{-1} = 1$. But since x^{-1} is the inverse element of x, it holds that $x^{-1} \cdot x = x \cdot x^{-1} = 1$, so it follows from Remark 1.26.(2) that $x = (x^{-1})^{-1}$.

b) We compute that

$$\begin{aligned}(x \cdot y) \cdot (y^{-1} \cdot x^{-1}) &\overset{\text{(G1)}}{=} ((x \cdot y) \cdot y^{-1}) \cdot x^{-1} \overset{\text{(G1)}}{=} (x \cdot (y \cdot y^{-1})) \cdot x^{-1} \\ &= (x \cdot e) \cdot x^{-1} \overset{\text{(G2)}}{=} x \cdot x^{-1} = e.\end{aligned}$$

Analogously, one shows that $(y^{-1} \cdot x^{-1}) \cdot (x \cdot y) = e$, so it follows from Remark 1.26.(2) that $(x \cdot y)^{-1} = y^{-1} \cdot x^{-1}$.

□

Theorem 1.28. *Let $(G, \cdot)$ be a group and $a, b \in G$. Then the equation $a \cdot x = b$ has the unique solution $x = a^{-1} \cdot b$.*

Proof. Obviously, $a^{-1} \cdot b$ is indeed a solution of $a \cdot x = b$. By (G3), a has an inverse element $a^{-1} \in G$. If we multiply both sided of the equation $b = a \cdot x$ by a^{-1}, we obtain

$$a^{-1} \cdot b = a^{-1} \cdot (a \cdot x) \overset{\text{(G1)}}{=} (a^{-1} \cdot a) \cdot x \overset{\text{(G3)}}{=} e \cdot x \overset{\text{(G2)}}{=} x.$$

□

As motivated in the beginning of this section, we next want to study sets with *two* algebraic operations that are in some sense "compatible" with each other, namely we want a *law of distributivity* to hold.

Definition 1.29. A *field* is a triple $(K, +, \cdot)$, where K is a set and

$$+ : K \times K \to K,$$
$$\cdot : K \times K \to K,$$

are maps satisfying the following properties:

(K1) $(K, +)$ is an abelian group,

(K2) $(K^*, \cdot)$ is an abelian group, where $K^* = K \setminus \{0\}$ and 0 denotes the neutral element of $(K, +)$.

(K3) *(distributivity)* For all $a, b, c \in K$ it holds that

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

We simply write K instead of $(K, +, \cdot)$ if it is clear which operations we are referring to. We always denote the neutral element of $+$ by 0 and the neutral element of $\cdot$ by 1 and call $+$ the *addition* and $\cdot$ the *multiplication* of K. The inverse element of $x \in K$ with respect to addition will be denoted by $-x$ and we put

$$x - y := x + (-y) \quad \forall x, y \in K.$$

We will further write $xy := x \cdot y$ for $x, y \in K$ and $\frac{x}{y} := x \cdot y^{-1}$ if $y \in K^*$.

Example 1.30. Let $\mathbb{Q} := \{\frac{p}{q} \mid p \in \mathbb{Z},\ q \in \mathbb{N}\}$ be the set of rational numbers, equipped with the usual rule for reduction and expansion of fractions, such that $\frac{a \cdot p}{a \cdot q} = \frac{p}{q}$ for all $a \in \mathbb{Z} \setminus \{0\}$. We define an addition and a multiplication on $\mathbb{Q}$ by

$$\cdot : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}, \quad \frac{a}{b} \cdot \frac{c}{d} := \frac{ac}{bd},$$
$$+ : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}, \quad \frac{a}{b} + \frac{c}{d} := \frac{ad + bc}{bd}.$$

Then $-\frac{p}{q} = \frac{(-p)}{q}$ and $(\frac{p}{q})^{-1} = \frac{q}{p}$ if $p \neq 0$.

Remark 1.31. (1) Note that, since $(K, +)$ is an abelian group, it particularly follows from Theorem 1.27 that

$$-(-x) = x \quad \text{and} \quad -(x + y) = -x - y \qquad \forall x, y \in K.$$

and from Theorem 1.28 that the equation

$$a + x = b$$

has the unique solution $x = b - a$ for all $a, b \in K$.

(2) Let $x^2 := x \cdot x$ for every $x \in K$. Using axiom (K3) one further observes that the following rules, called "binomische Formeln" in German high schools, hold for all $a, b \in K$:

$$(a + b)^2 = a^2 + 2ab + b^2,$$
$$(a - b)^2 = a^2 - 2ab + b^2,$$
$$(a + b)(a - b) = a^2 - b^2.$$

We can derive some elementary properties of fields straight from their definition.

Theorem 1.32. *Let $(K, +, \cdot)$ be a field. For all $x \in K$ it holds that $x \cdot 0 = 0$.*

Proof. Let $x \in K$. Since 0 is the neutral element of $+$, it holds that

$$x \cdot 0 = x \cdot (0 + 0) \overset{(K3)}{=} x \cdot 0 + x \cdot 0.$$

Adding $-(x \cdot 0)$ to both sides of the equation yields

$$0 = (x \cdot 0 + x \cdot 0) + (-x \cdot 0) = x \cdot 0 + (x \cdot 0 + (-x \cdot 0)) = x \cdot 0,$$

what we wanted to show. □

Theorem 1.33. *If $(K, +, \cdot)$ is a field and $a, b \in K$, then the following equations hold:*

a) $(-a) \cdot b = -(a \cdot b)$.

b) $(-a) \cdot (-b) = a \cdot b$.

Proof. This is exercise 3 on Sheet 1. □

Theorem 1.34. *Let K be a field and $a, b \in K$. If $a \cdot b = 0$, then $a = 0$ or $b = 0$.*

Proof. Let $a, b \in K$ with $a \cdot b = 0$ and assume that $a \neq 0$, such that $a^{-1} \in K$ is well-defined. Then by Theorem 1.32,

$$0 = a^{-1} \cdot 0 = a^{-1} \cdot (a \cdot b) = (a^{-1} \cdot a) \cdot b = b.$$

□

There is one more crucial property of the real numbers that we want to formalize in this section. Geometrically, the real numbers are imagined as an infinite line that has a direction with numbers getting bigger and bigger if one follows this direction. To put it differently, the real numbers are *ordered*. Given two real numbers $x, y \in \mathbb{R}$ one can always tell whether $x \leq y$ holds or not. This notion of ordering can be defined in abstract terms for fields as well. We start with a set-theoretic notion of "comparing" elements of sets with each other.

Definition 1.35. Let M be a set.

a) A *binary relation on M* is a subset $R \subset M \times M$. For a binary relation R we write xRy if $(x, y) \in R$.

b) Let M be a set. A *total order on M* is a binary relation $\leq$ on M with the following properties:

   (i) $x \leq x$ for every $x \in M$. *(reflexivity)*

   (ii) If $x, y \in M$ satisfy $x \leq y$ and $y \leq x$, then $x = y$. *(antisymmetry)*

   (iii) If $x, y, z \in M$ satisfy $x \leq y$ and $y \leq z$, then $x \leq z$. *(transitivity)*

   (iv) For all $x, y \in M$ it holds that $x \leq y$ or $y \leq x$. *(totality)*

Considering fields, we want our order to behave well with respect to addition and multiplication.

Definition 1.36. a) A field $(K, +, \cdot)$ is called *ordered* if it admits a total order $\leq$ with the following properties:

(O1) If $x, y \in K$ with $x \leq y$, then
$$x + z \leq y + z \quad \forall z \in K.$$

(O2) If $x, y, z \in K$ satisfy $x \leq y$ and $0 \leq z$, then $xz \leq yz$.

b) If $(K, +, \cdot)$ is an ordered field, then $x \in K$ is called

- *non-negative* if $0 \leq x$,
- *positive* if $0 \leq x$ and $x \neq 0$,
- *non-positive* if $x \leq 0$,
- *negative* if $x \leq 0$ and $x \neq 0$.

For $x, y \in K$ we further write

- $x < y$ if and only if $x \leq y$ and $x \neq y$,
- $x \geq y$ if and only if $y \leq x$,
- $x > y$ if and only if $y < x$.

Example 1.37. $(\mathbb{Q}, +, \cdot)$ is indeed an ordered field with respect to the usual relation $\leq$, i.e. that $\frac{a}{b} \leq \frac{c}{d}$ if and only if $ad \leq bc$ holds in $\mathbb{Z}$.

In the following we will denote

$$x^n := \underbrace{x \cdot x \cdot \ldots \cdot x}_{n \text{ times}}, \qquad x^{-n} := (x^{-1})^n = \underbrace{x^{-1} \cdot x^{-1} \cdot \ldots \cdot x^{-1}}_{n \text{ times}},$$
$$nx = \underbrace{x + x + \ldots + x}_{n \text{ times}}, \qquad -nx = n(-x) = \underbrace{-x - x - \ldots - x}_{n \text{ times}},$$

for all $n \in \mathbb{N}$ and $x \in K$. We further put $x^0 := 1$.
The following theorem collects important inequalities for ordered fields that can be immediately derived from their defining axioms and our previous results about fields. Therefore, we omit its proof. While these properties might look confusing upon first glance, they actually ensure that many of the computational rules for real numbers that we have in mind from high school actually hold true in general ordered fields.

Theorem 1.38. *Let K be an ordered field.*

a) For all $a, b, c \in K$ with $a < b$ it holds that $a + c < b + c$. If additionally $c > 0$, then $a \cdot c < b \cdot c$.

b) If $a, b, c \in K$ satisfy $a \leq b$ and $b < c$, then $a < c$.

c) For all $a, b, c, d \in K$ with $a \leq b$ and $c \leq d$ it holds that
$$a + c \leq b + d.$$
If additionally $c < d$, then $a + c < b + d$.

d) If $a \in K$ with $a \geq 0$, then $-a \leq 0$. If $a > 0$, then $-a < 0$.

e) For all $a, b, c \in K$ with $a \leq b$ and $c \leq 0$, it holds that
$$ac \geq bc.$$
If additionally $a < b$ and $c < 0$, then $ac > bc$.

f) For all $a \in K$ it holds that $a^2 \geq 0$. If $a \neq 0$, then $a^2 > 0$.

g) If $a \in K$ satisfies $a > 0$, then $a^{-1} > 0$.

h) If $a, b \in K$ satisfy $a \leq b$, then $a^{-1} \geq b^{-1}$. If $a < b$, then $a^{-1} > b^{-1}$.

The notion of absolute value of a real number can now be transferred to arbitrary ordered fields.

Definition 1.39. Let K be an ordered field and let $x \in K$. The *absolute value of x* is given by

$$|x| := \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Theorem 1.40. *Let K be an ordered field. For all $x, y \in K$ the following properties hold:*

a) $|x| \geq 0$.

b) $|x| = 0$ if and only if $x = 0$.

c) $|x \cdot y| = |x| \cdot |y|$.

d) (triangle inequality) $|x + y| \leq |x| + |y|$.

Proof. a) If $x \geq 0$, the claim is obvious. If $x < 0$, then $-x > 0$ by Theorem 1.38.d). Hence $|x| = -x > 0$.

b) Since trivially $0 \geq 0$, it holds by definition that $|0| = 0$. Conversely, if $x \in K$ satisfies $|x| = 0$, then by definition of the absolute value, it needs to hold that $x \geq 0$ such that $x = |x| = 0$.

c) If $x = 0$ or $y = 0$, the statement follows from Theorem 1.32.

If $x > 0$ and $y > 0$, the statement is obvious.

If $x > 0$ and $y < 0$, then $x \cdot y < 0$ by Theorem 1.38.e), hence

$$|x \cdot y| = -x \cdot y = x \cdot (-y) = |x| \cdot |y|,$$

where we used Theorem 1.33.a). The case $x < 0$ and $y > 0$ follows analogously.

If $x < 0$ and $y < 0$, then $xy > 0$ by Theorem 1.38.e). Using Theorem 1.33.b) we obtain

$$|x \cdot y| = x \cdot y = (-x) \cdot (-y) = |x| \cdot |y|,$$

which completes the proof.

d) By definition of $|\cdot|$, it holds for every $a \in K$ that $a \leq |a|$ and $-a \leq |a|$. Using Theorem 1.38.c) it follows that

$$\begin{aligned} x + y &\leq |x| + |y|, \\ -(x + y) = -x - y &\leq |x| + |y|. \end{aligned}$$

Since $|x + y| \in \{x + y, -(x + y)\}$, we derive that $|x + y| \leq |x| + |y|$.

□

We want the real numbers to fulfill the axioms of an ordered field. However, the real numbers are not the only ordered field that suffice these axioms as the following example shows.
It turns out that we need an additional axiom to determine the real numbers that we will introduce in the next section. So far, we observe that the rational numbers are "lacking" some numbers in the following sense: since $\mathbb{Q}$ is an ordered field, it holds by Theorem 1.38.d) that $x^2 \geq 0$ for every $x \in \mathbb{Q}$. However, there are numbers $a \in \mathbb{Q}$ with $a \geq 0$, such that $x^2 = a$ is not fulfilled for any $x \in \mathbb{Q}$. For example, one can show that $x^2 = 2$ does not have any solution in $\mathbb{Q}$. We will see that this problem does not occur anymore if we demand our ordered field to satisfy one additional condition.

1.3 Real numbers

To make the "flaws" of $\mathbb{Q}$ more precise, we will introduce bounded subsets of ordered fields and formalize the property that we want the real numbers to have in order to ensure the existence of solutions of certain equations. This property allows us to uniquely determine the real numbers as an ordered field satisfying an additional axiom.

Definition 1.41. Let K be an ordered field and $A \subset K$ be a subset.

a) A is called *bounded from above* if there exists $c \in K$, such that $a \leq c$ for all $a \in A$. c is then called *an upper bound of* A.

b) A is called *bounded from below* if there exists $b \in K$, such that $a \geq b$ for all $a \in A$. b is then called *a lower bound of* A.

c) $s \in K$ is called *the smallest upper bound of* A if s is an upper bound of A and if every other upper bound c of A satisfies $s \leq c$.

d) $i \in K$ is called *the biggest lower bound of* A if i is a lower bound of A and if every other lower bound b of A satisfies $b \leq i$.

Remark 1.42. An equivalent way of describing smallest upper bounds and biggest lower bounds in ordered fields is the following.

(1) s^* is the smallest upper bound of A if and only if s^* is an upper bound of A and

$$\forall s < s^* \; \exists x \in A \;\text{ with }\; x > s.$$

(2) i^* is the biggest lower bound of A if and only if i^* is a lower bound of A and

$$\forall i > i^* \; \exists x \in A \;\text{ with }\; x < i.$$

The existence of smallest upper and biggest lower bounds is in general not guaranteed. For example, in the ordered field $K = \mathbb{Q}$ it can be shown that the set

$$\{x \in \mathbb{Q} \mid x^2 < 2\}$$

has no smallest upper bound, although it is bounded from above.

Definition 1.43. An ordered field K is *complete* if every subset of A that is bounded from above possesses a smallest upper bound.

This assumption turns out to be the missing piece for our definition of real numbers.

Theorem/Definition 1.44. *Up to "equivalence"*[4]*, there exists a unique complete ordered field. We call this field* the real numbers *and denote it by* $\mathbb{R}$.

We will omit the proof of this result and proceed by further studying the real numbers.

Theorem 1.45. *Every* $A \subset \mathbb{R}$ *that is bounded from below possesses a biggest lower bound.*

Proof. Put $-A := \{-x \in \mathbb{R} \mid x \in A\}$. If $b \in \mathbb{R}$ is a lower bound of A, i.e. $b \leq x$ for all $x \in A$, then it follows from Theorem 1.38.e) that

$$-x \leq -b \qquad \forall x \in A.$$

Hence, $-A$ is bounded from above by $-b$ and by construction of $\mathbb{R}$, $-A$ has a smallest upper bound $s \in \mathbb{R}$. Then $-x \leq s$ for all $x \in A$, such that

$$x \geq -s \qquad \forall x \in A,$$

i.e. $-s$ is a lower bound of A. Suppose there was an $a \in \mathbb{R}$ with $a > -s$ which is a lower bound of A. Then $x \geq a$ for all $x \in A$, which implies $-x \leq -a$ for all $x \in A$, i.e. $-a$ is an upper bound of $-A$. Since $a > -s$, it further holds that $-a < s$. ↯[5]
This is a contradiction, since s is the *smallest* upper bound of $-A$, see Remark 1.42.a). □

Definition 1.46. Let $A \subset \mathbb{R}$.

a) If A is bounded from above, we let $\sup A$ denote the smallest upper bound of A and call it *the supremum of* A. If A is not bounded from above, we put $\sup A := +\infty$.

b) If A is bounded from above and $s := \sup A$ satisfies $s \in A$, then s is called *the maximum of* A and denoted by $\max A$.

c) If A is bounded from below, we let $\inf A$ denote the biggest lower bound of A and call it *the infimum of* A. If A is not bounded from below, we put $\inf A := -\infty$.

d) If A is bounded from below and $i := \inf A$ satisfies $i \in A$, then i is called *the minimum of* A and denoted by $\min A$.

Example 1.47. Let $I := \{x \in \mathbb{R} \mid 0 \leq x < 2\}$. Then every $C \geq 2$ is an upper bound of I while every $c \leq 0$ is a lower bound of I. Since 0 is a lower bound of I and since $0 \in I$, it follows that $\inf I = \min I = 0$. On the other hand, 2 is an upper bound of I and one can show (see Theorem 1.56 below) that for every $C \in \mathbb{R}$ with $C < 2$ there exists an $x \in I$ with $C < x < 2$, so no such C can be an upper bound of I. Thus, $\sup I = 2$. Since $2 \notin I$, the set I possesses no maximum.

An important consequence of the completeness of $\mathbb{R}$ is the following theorem, which could be proven with a lot of effort directly from our construction, but which we will prove later in this course by a simpler argument.

Theorem 1.48. *Let* $a \in \mathbb{R}$ *with* $a \geq 0$ *and let* $n \in \mathbb{N}$. *There exists a unique* $x \in \mathbb{R}$ *with* $x \geq 0$, *such that* $x^n = a$.

[4] For the curious reader: The mathematically correct notion would be *up to an isomorphism of fields*.
[5] A lightning symbol is often used in mathematical proofs to mark a contradiction.

Definition 1.49. Let $a \in \mathbb{R}$ with $a \geq 0$ and let $n \in \mathbb{N}$. The unique solution of $x^n = a$ with $x \geq 0$ is called *the n-th root of a* and denoted by $\sqrt[n]{a}$ or $a^{\frac{1}{n}}$. We also put $\sqrt{a} := \sqrt[2]{a}$ and call it *the square root of a*. Additionally, for each $\frac{p}{q} \in \mathbb{Q}$ we define

$$a^{\frac{p}{q}} := \left(a^{\frac{1}{q}}\right)^p = \left(\sqrt[q]{a}\right)^p.$$

The sets of numbers $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$, that we have informally introduced before, can now be redefined as subsets of $\mathbb{R}$.

Definition 1.50. a) We recall that $1 \in \mathbb{R}$ denotes the neutral element of multiplication in $\mathbb{R}$. A subset $M \subset \mathbb{R}$ is called *inductive* if

(i) $1 \in M$,

(ii) If $x \in \mathbb{R}$ satisfies $x \in M$, then $x + 1 \in M$.

b) We define the *natural numbers* $\mathbb{N} \subset \mathbb{R}$ by

$$\mathbb{N} := \{x \in \mathbb{R} \mid x \in M \text{ for all inductive subsets } M \subset \mathbb{R}\}.$$

One checks without difficulties that $\mathbb{N}$ is itself inductive.

Remark 1.51. One checks from the definition that $\mathbb{N}$ is given by

$$\mathbb{N} = \{1, 1+1, (1+1)+1, ((1+1)+1)+1, \dots\} =: \{1, 2, 3, 4, \dots\}.$$

Theorem 1.52. $\mathbb{N}$ *is not bounded from above.*

Proof. Suppose $\mathbb{N}$ was bounded from above and put $s := \sup \mathbb{N} < +\infty$. Since s is the *lowest* upper bound of $\mathbb{N}$, the number $s - 1 \in \mathbb{R}$ is not an upper bound for $\mathbb{N}$. Hence, there exists $n \in \mathbb{N}$ with $s - 1 < n$, which implies $s < n + 1$ by Theorem 1.38.c). But since $\mathbb{N}$ is inductive, it holds that $n + 1 \in \mathbb{N}$, which contradicts the fact that s is an upper bound for $\mathbb{N}$. Thus, $\mathbb{N}$ is not bounded from above. □

Corollary 1.53. *For every $\varepsilon > 0$ there exists an $n \in \mathbb{N}$ with $\frac{1}{n} < \varepsilon$.*

Proof. Since $\mathbb{N}$ is not bounded from above, there exists an $n \in \mathbb{N}$ with $n > \frac{1}{\varepsilon}$. By Theorem 1.38.h) this yields the claim. □

Now that we have formally defined $\mathbb{N}$, the formal definitions of $\mathbb{Z}$ and $\mathbb{Q}$ are very straightforward.

Definition 1.54. a) The *integers* are as a subset of $\mathbb{R}$ defined by

$$\mathbb{Z} := \mathbb{N} \cup \{0\} \cup \{x \in \mathbb{R} \mid -x \in \mathbb{N}\}.$$

We further let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ denote the set of non-negative integers.

b) The *rational numbers* are as a subset of $\mathbb{R}$ defined by

$$\mathbb{Q} := \{p \cdot q^{-1} \in \mathbb{R} \mid p \in \mathbb{Z},\ q \in \mathbb{N}\}.$$

Clearly, $\mathbb{Z} \subset \mathbb{Q}$.

We conclude this section by studying the behaviour of the rational numbers inside the real numbers.

Definition 1.55. Given $x \in \mathbb{R}$ we let $\lfloor x \rfloor$ denote the biggest integer which is not bigger than x, i.e.

$$\lfloor x \rfloor = \sup\{n \in \mathbb{Z} \mid n \leq x\}.$$

Obviously, $\lfloor x \rfloor \leq x$. (Some authors denote $\lfloor x \rfloor$ by $[x]$ instead.)

Theorem 1.56. *Let $x, y \in \mathbb{R}$ with $x < y$. Then there exists $q \in \mathbb{Q}$, such that $x < q < y$.*

Proof. If $\lfloor x \rfloor + 1 < y$, then put $q := \lfloor x \rfloor + 1 > x$.
Assume now that $x < y < \lfloor x \rfloor + 1$. By Corollary 1.53, there is an $n \in \mathbb{N}$, such that

$$\frac{1}{n} < y - x. \tag{1.1}$$

Let $m_0 := \sup\{k \in \mathbb{Z} \mid \lfloor x \rfloor + \frac{k}{n} \leq x\} + 1$ and put $m := m_0 + 1$ and $q := \lfloor x \rfloor + \frac{m}{n} \in \mathbb{Q}$. Then, by definition of m, it holds that $x < q$. We further compute that

$$q = \lfloor x \rfloor + \frac{m}{n} = \lfloor x \rfloor + \frac{m_0}{n} + \frac{1}{n} \leq x + \frac{1}{n} \overset{(1.1)}{<} y.$$

Hence, q has the desired properties. □

1.4 Mathematical induction

This section's aim is to introduce a new proof technique, the so-called mathematical induction. So far, we have seen *direct proofs* and *proofs by contradiction*, now we will add a third method, the so-called *proof by induction*. This method is used to prove statements that depend on some parameter n, where $n \in \mathbb{N}$ might be arbitrary. The following theorem lays the foundation for this technique.

Theorem 1.57 (The induction principle)**.** *Let $M \subset \mathbb{N}$ be a subset with the following properties:*

(i) $1 \in M$,

(ii) For each $n \in \mathbb{N}$ with $n \in M$ it holds that $n + 1 \in M$.

Then $M = \mathbb{N}$.

Proof. It follows from properties (i) and (ii) that M is an inductive set. Since $\mathbb{N}$ is by definition contained in *every* inductive subset of $\mathbb{R}$, it follows that $\mathbb{N} \subset M$. Since the other inclusion holds by assumption, it follows that $M = \mathbb{N}$. □

The reader might wonder about the relevance of the induction principle for our considerations. Its key consequence is the following corollary.

Corollary 1.58. *Let $A(n)$ be a logical proposition making a statement about $n \in \mathbb{N}$. If the following properties hold:*

(i) $A(1)$ is true,

(ii) if $A(n)$ is true for some $n \in \mathbb{N}$, then $A(n+1)$ is true,

then $A(n)$ is true for every $n \in \mathbb{N}$.

Proof. This follows from applying the induction principle to the set $M = \{n \in \mathbb{N} \mid A(n) \text{ is true}\}$. □

Corollary 1.58 provides us with a powerful method to prove the validity of a statement, e.g. an equation, about natural numbers. If $A(n)$ is such a statement, we need to go through the following steps:

- Prove the claim explicitly for $n = 1$. This step is called the *base case*.
- Assume that $A(n)$ is true for an arbirtrary $n \in \mathbb{N}$. This assumption is called the *induction hypothesis*.
- Prove that, under the induction hypothesis, i.e. assumption that $A(n)$ is true, it follows that $A(n+1)$ is true as well. This step is called the *induction step*.

By Corollary 1.58, it follows from these steps that $A(n)$ is true for *all* $n \in \mathbb{N}$.
In the remainder of this section, we will discuss several important results which are proven by mathematical induction. We start by summarizing some elementary properties of $\mathbb{N}$ as we defined it in the previous section and will only prove a part of it. The reader is encouraged to fill in proofs of the remaining parts of the theorem.

Theorem 1.59. *Let $m, n \in \mathbb{N}$. Then:*

a) $m + n \in \mathbb{N}$,

b) $m \cdot n \in \mathbb{N}$.

c) $n \geq 1$.

d) if $m < n$, then $n - m \in \mathbb{N}$.

Proof of part a) by induction. We consider $m \in \mathbb{N}$ as fixed and will prove the result by induction over n.

Base case: $n = 1$
Since $m \in \mathbb{N}$ and $\mathbb{N}$ is an inductive set, it holds that $m + 1 \in \mathbb{N}$. ✓

Induction step: Assume that $m + n \in \mathbb{N}$ for some given $n \in \mathbb{N}$. From the associativity of addition in $\mathbb{R}$, we derive:

$$m + (n + 1) = (m + n) + 1.$$

Since $m + n \in \mathbb{N}$ by our induction hypothesis and since $\mathbb{N}$ is inductive, it follows that $(m + n) + 1 \in \mathbb{N}$, which shows that $m + (n + 1) \in \mathbb{N}$ and thereby concludes the induction step.
Thus, we have shown that $m + n \in \mathbb{N}$ for all $n \in \mathbb{N}$. □

Before we prove three very important results, we first introduce some additional shorthand notation.

Definition 1.60. a) Let K be a field and $a_1, \ldots, a_n \in K$. We define

$$\sum_{i=1}^{n} a_i := a_1 + a_2 + \cdots + a_n \in K.$$

b) For each $n \in \mathbb{N}$ we define $n!$ (spoken: "n factorial") as

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

We further put $0! := 1$.

Example 1.61. If we want to sum up all odd numbers up to $2n-1$ for some $n \in \mathbb{N}$, this can be written as

$$1+3+5+7+9+\cdots+(2n-3)+(2n-1) = \sum_{i=1}^{n+1}(2i-1).$$

The following result was allegedly proven by the famous mathematician Carl Friedrich Gauß when he was nine (!) years old and is sometimes called "the little Gauß" because of his fact. Since young Carl Friedrich was better in math than the rest of his class, his teacher gave him the task of summing up all numbers from 1 to 100 to keep him busy and quiet for a while. When young Carl Friedrich came up with the answer after a very short amount of time because he had developed a nice formula for it, his teacher was stunned. (At least this is the way the story goes...)

Theorem 1.62 (Gauß)**.** *For each $n \in \mathbb{N}$ it holds that*

$$1+2+\cdots+n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Proof. We will prove the claim by induction over n.
Base case: $n = 1$

We compute both sides of the equation explicitly. The left-hand side is $\sum_{i=1}^{1} i = 1$. The right-hand side is given by

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

This shows the base case. ✓

Induction step: We need to show that, under the assumption that $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ for some $n \in \mathbb{N}$, it follows that

$$\sum_{i=1}^{n+1} i \overset{!}{=} \frac{(n+1)(n+2)}{2}.$$

We compute that

$$\sum_{i=1}^{n+1} i = \sum_{i=1}^{n} i + n + 1 \overset{(*)}{=} \frac{n(n+1)}{2} + n + 1 = \frac{n^2+n+2(n+1)}{2} = \frac{n^2+3n+2}{2} = \frac{(n+1)(n+2)}{2},$$

where we have used the induction hypothesis in $(*)$. This completes the induction step and thus the proof. □

Theorem 1.63 (Bernoulli's inequality)**.** *Let $x \in \mathbb{R}$ with $x \geq -1$ and $n \in \mathbb{N}$. Then*

$$(1+x)^n \geq 1 + nx.$$

Proof. We prove the claim by induction over n and consider $x \geq -1$ as fixed.

Base case: $n = 1$
This case is obvious, since $(1+x)^1 = 1 + x = 1 + 1 \cdot x$. ✓

Induction step: We want to show that $(1+x)^{n+1} \overset{!}{\geq} 1 + (n+1)x$ under the assumption that $(1+x)^n \geq 1 + nx$. Since $x \geq -1$, it holds that $1 + x \geq 0$. Thus, using ordering axiom (O2), we compute that

$$\begin{aligned}(1+x)^{n+1} = (1+x)^n \cdot (1+x) &\overset{(*)}{\geq} (1+nx)(1+x) \\ &= 1 + nx + x + x^2 \\ &= 1 + (n+1)x + x^2 \geq 1 + (n+1)x\end{aligned}$$

by ordering axiom (O2) and Theorem 1.38.f), where we have used the induction hypothesis in $(*)$. □

The following result generalizes the formula from Remark 1.31.(2) to higher exponents.

Theorem 1.64 (Binomial theorem)**.** *For all $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$, it holds that*

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

where

$$\binom{n}{k} := \frac{n!}{k! \cdot (n-k)!} \qquad \forall k \in \{0, 1, \ldots, n\}.$$

*Proof. (This proof is **not** given in the lecture.)*

Fix numbers $a, b \in \mathbb{R}$.

Base case: $n = 1$
One easily checks that $\binom{1}{0} = 1$ and $\binom{1}{1} = 1$, so we obtain

$$(a+b)^1 = a + b = \binom{1}{1} a^1 b^0 + \binom{1}{0} a^0 b^1 = \sum_{k=0}^{1} a^k b^{1-k},$$

which we wanted to show. ✓

Induction step: We want to show that

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

under the assumption that $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$ for some $n \in \mathbb{N}$. We first compute:

$$\begin{aligned}(a+b)^{n+1} &= (a+b)^n \cdot (a+b) \\ &\overset{(*)}{=} \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \cdot (a+b) \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{k+1} b^{n-k} + \sum_{k=0}^{n} \binom{n}{k} a^k b^{n+1-k}.\end{aligned}$$

where we used the distributivity of $\mathbb{R}$ and, in $(*)$, the induction hypothesis. Now in the first of the two sums we perform a so-called *index shift*, we simply add 1 to the variable k and transform the formula and the summation bounds accordingly.

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

We want to carry out some computations regarding the coefficients involved. We first observe that

$$\binom{n}{n} = \frac{n!}{0! \cdot n!} = \frac{n!}{n!} = 1, \qquad \binom{n}{0} = \frac{n!}{n! \cdot 0!} = \frac{n!}{n!} = 1.$$

For the coefficients inside the sum we compute for each $k \in \{1, 2, \ldots, n\}$:

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

Since

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

this yields

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

Inserting these computations into the above computation shows that

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

This is what we wanted to show. □

Remark 1.65. The numbers $\binom{n}{k}$ (spoken "n choose k") are called *binomial coefficients* and motivated by elementary combinatorics: if one has a set with n elements and wants to choose k different elements of this set, there are precisely $\binom{n}{k}$ possibilities of doing so. For example, in the "Lotto", the biggest German lottery system, there is a bowl containing 49 balls labelled from

1 to 49. Each week six of this balls are randomly chosen by a machine and the participants have to bet their money on six numbers, i.e. on a fixed subset of $\{1, 2, \ldots, 49\}$ containing six elements. For doing so, there are $\binom{49}{6} = \frac{49!}{43! \cdot 6!} = 13983816$ possibilities. So the chances of winning the jackpot are very, very, very slim...

1.5 Complex numbers

While we have seen that $x^n = a$ has a solution in $\mathbb{R}$ whenever $a \geq 0$, in general there might not be a real solution of $x^n = a$ if $a < 0$. For example, there is no real solution of

$$x^2 = -1.$$

For certain computations one would like to extend the field of real numbers to a bigger field K in which $x^n = a$ has solutions for all $a \in K$. It turns out that the field $K = \mathbb{C}$ of *complex numbers* is the right extension for these purposes. In this section, we will give a first introduction to complex numbers and their properties and will pick them up later in this lecture to study them further.

We define an addition and a multiplication on $\mathbb{R}^2$ by

$$\begin{aligned}(x_1, y_1) + (x_2, y_2) &= (x_1 + x_2, y_1 + y_2),\\ (x_1, y_1) \cdot (x_2, y_2) &= (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1) \qquad \forall (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2.\end{aligned}$$

We consider

$$\mathbb{C} := \{x + iy \mid (x, y) \in \mathbb{R}^2\},$$

where the i is a purely formal symbol. It is clear that the map $\Phi : \mathbb{C} \to \mathbb{R}^2$, $\Phi(x + iy) = (x, y)$ is a bijection. The addition and multiplication from above can be seen as maps on $\mathbb{C}$, where they read as follows:

$$\begin{aligned}(x_1 + iy_1) + (x_2 + iy_2) &= x_1 + x_2 + i(y_1 + y_2),\\ (x_1 + iy_1) \cdot (x_2 + iy_2) &= x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1) \qquad \forall x_1 + iy_1, x_2 + iy_2 \in \mathbb{C}.\end{aligned}$$

Theorem/Definition 1.66. *$(\mathbb{C}, +, \cdot)$ is a field, for which:*

- *the neutral element of $+$ is $0 = 0 + i \cdot 0$,*
- *the neutral element of $\cdot$ is $1 = 1 + i \cdot 0$,*
- *$-(x + iy) = -x + i(-y)$ for all $x + iy \in \mathbb{C}$,*
- *$(x + iy)^{-1} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} - i\frac{y}{x^2 + y^2}$ for all $x + iy \in \mathbb{C}^*$.*

$(\mathbb{C}, +, \cdot)$ is called the field of complex numbers.

Proof. One checks by direct computations (exercise!) that $(\mathbb{C}, +)$ and $(\mathbb{C}^*, \cdot)$ are abelian groups, that the law of distirbutivity (K3) holds and that the additive inverse is of the claimed form. We check the formula for $(x + iy)^{-1}$ by another direct computation:

$$\begin{aligned}(x + iy)(x + iy)^{-1} &= (x + iy)\Big(\frac{x}{x^2 + y^2} - i\frac{y}{x^2 + y^2}\Big)\\ &= \frac{x^2}{x^2 + y^2} - \frac{-y^2}{x^2 + y^2} + i \cdot \Big(-\frac{xy}{x^2 + y^2} + \frac{xy}{x^2 + y^2}\Big) = 1 + i \cdot 0 = 1.\end{aligned}$$

□

Remark 1.67. (1) One checks from their definitions that the restrictions of $+$ and $\cdot$ to $\{x + i \cdot 0 \mid x \in \mathbb{R}\}$ correspond to addition and multiplication in $\mathbb{R}$. Thus, we can identify $x \in \mathbb{R}$ with $x + i \cdot 0$ and view $\mathbb{R}$ as a subset of $\mathbb{C}$ with the same addition and multiplication. Formally, one says that $\mathbb{R}$ is a *subfield* of $\mathbb{C}$ in this situation.

(2) $1 = 1 + i \cdot 0$ is called the *real unit* of $\mathbb{C}$, while $i = 0 + i \cdot 1$ is called the *imaginary unit* of $\mathbb{C}$. Note that

$$i^2 = (0 + i \cdot 1)(0 + i \cdot 1) = -1 + i \cdot 0 = -1.$$

Definition 1.68. Let $z = x + iy \in \mathbb{C}$.

a) $\bar{z} := x - iy$ is the *complex conjugate of* z.

b) $\mathrm{Re}(z) := x$ is called the *real part of* z and $\mathrm{Im}(z) := y$ is called the *imaginary part of* z.

Remark 1.69. Geometrically, viewing $\mathbb{C}$ as the plane $\mathbb{R}^2$, the complex conjugate of z is the reflection of z along the x-axis, i.e. the line $\mathbb{R} \times \{0\} \subset \mathbb{R}^2$.

We summarize some computational rules for complex conjugates in the following theorem, whose proof involves only straightforward computations and is strongly recommended as an exercise to get familiar with complex numbers.

Theorem 1.70. *Let $z, w \in \mathbb{C}$. Then:*

a) $\overline{z + w} = \bar{z} + \bar{w}$ *and* $\overline{zw} = \bar{z} \cdot \bar{w}$.

b) $\bar{\bar{z}} = z$, $z + \bar{z} = 2\mathrm{Re}(z)$ *and* $z - \bar{z} = 2i\mathrm{Im}(z)$.

c) $z = \bar{z}$ *if and only if* $z \in \mathbb{R}$.

d) If $z = x + iy$, *then* $z\bar{z} = x^2 + y^2$.

Definition 1.71. The *absolute value* of a complex number $z = x + iy \in \mathbb{C}$ is defined by

$$|z| := \sqrt{x^2 + y^2} = \sqrt{z\bar{z}}.$$

Remark 1.72. (1) Identifying $\mathbb{C}$ with the Euclidean plane $\mathbb{R}^2$, the absolute value of z corresponds to the length (Euclidean norm) of z as a vector in $\mathbb{R}^2$.

(2) If $x \in \mathbb{R} \subset \mathbb{C}$, then its absolute value is $\sqrt{x^2}$ which is given by $|x|$, the absolute value from Section 1.2. Thus, the "new" absolute value coincides with the "old" one on $\mathbb{R} \subset \mathbb{C}$.

The following result collects several properties of the absolute value of complex numbers.

Theorem 1.73. *For all $z, w \in \mathbb{C}$, the following properties hold:*

a) $|z| \geq 0$ *with* $|z| = 0$ *if and only if* $z = 0$.

b) $|\bar{z}| = |z|$.

c) $|\mathrm{Re}(z)| \leq |z|$ *and* $|\mathrm{Im}(z)| \leq |z|$.

d) $|zw| = |z| \cdot |w|$.

e) (triangle inequality) $|z + w| \leq |z| + |w|$.

Proof. Parts a), b) and c) follow immediately from the definition. For part d), we compute using Theorem 1.70.a) that

$$|zw|^2 = zw\overline{zw} = zw\bar{z}\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2.$$

By a), this implies $|zw| = |z| \cdot |w|$.
To prove e), we first compute using Theorem 1.70 that

$$\begin{aligned}
|z+w|^2 &= (z+w)(\bar{z}+\bar{w}) = z\bar{z} + z\bar{w} + w\bar{z} + w\bar{w} \\
&= |z|^2 + z\bar{w} + \overline{z\bar{w}} + |w|^2 \\
&= |z|^2 + 2\mathrm{Re}(z\bar{w}) + |w|^2 \\
&\leq |z|^2 + 2|z\bar{w}| + |w|^2 \\
&= |z|^2 + 2|z| \cdot |w| + |w|^2 = (|z| + |w|)^2.
\end{aligned}$$

Taking the square root yields $|z+w| \leq |z| + |w|$. □

As the next theorem shows, there is one fundamental difference between the real and the complex numbers.

Theorem 1.74. *There exists no total order on $\mathbb{C}$ which makes it an ordered field.*

Proof. Suppose that $\mathbb{C}$ was an ordered field with respect to a total order $\leq$. Then it particularly holds that $i \geq 0$ or $0 \leq i$. If $i \geq 0$, then $i^3 \geq 0$ as well by Theorem 1.38, but $i^3 = i^2 \cdot i = -i$. Hence, $-i \geq 0$ and thus $i \leq 0$ which is a contradiction to $i \geq 0$ and $i \neq 0$.
If $i \leq 0$, then $-i \geq 0$ and thus $(-i)^3 \geq 0$. However, $(-i)^3 = (-1)^3 i^3 = -(-i) = i \leq 0$, which is again a contradiction. Thus, there is no structure of an ordered field on $\mathbb{C}$. □

We will investigate the complex numbers further at a later point in this lecture. So far, we want to conclude this section by stating a famous result showing the importance and the most fundamental property of complex numbers. Unfortunately, we have not developed enough tools to prove it, so the reader might just take it as a motivation to get interested in $\mathbb{C}$ at this point.

Theorem 1.75 (Fundamental theorem of algebra)**.** *For all $n \in \mathbb{N}$ and $a_0, a_1, \ldots, a_{n-1} \in \mathbb{C}$ the equation*

$$z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 = 0$$

has a solution in $\mathbb{C}$.

Chapter 2

Sequences and series

2.1 Sequences and convergence

Intuitively, a sequence is an infinite list of items taken from a particular set, e.g. an infinite list of numbers. Like any good list, a sequence has a start and an ordering of items. Sequences occur everywhere in our daily lives. For example, a movie or TV show being watched on a TV is nothing but a sequence of still images - presented to us so quick that we don't notice that we are actually scrolling through pictures at an enormous pace.
Putting a cup of hot tea into a room of regular temperature and a thermometer into the cup, one might simply make a list of real numbers by measuring the temperature of the tea once every minute, so the n-th item would be the temperature after n minutes. As we know from our experience and as you will learn in thermodynamics, the temperature of the tea will go down and *approach* or *tend to* the temperature of the air it is surrounded by. This phenomenon of *approaching* something is precisely what is formalized by the notion of convergence that we will see in this section. First of all, we will formalize the notion of an "infinite list".

Definition 2.1. Let X be a set. A *sequence in* X is a map $a : \mathbb{N} \to X$. We denote such a sequence also by

$$(a_n)_{n\in\mathbb{N}}, \quad \text{where} \quad a_n := a(n) \ \text{ for all } \ n \in \mathbb{N},$$

or by $(a_1, a_2, a_3, \dots)$. The a_n are called the *elements* or *terms* of the sequence. A sequence in $\mathbb{R}$ is called a *real sequence*, while a sequence in $\mathbb{C}$ is called a *complex sequence*.

Example 2.2. Real sequences:

a) $a_n = \frac{1}{n}, \quad (a_n)_{n\in\mathbb{N}} = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots)$.

b) $b_n = (-1)^n, \quad (b_n)_{n\in\mathbb{N}} = (-1, 1, -1, 1, \dots)$.

c) $c_n = n^2, \quad (c_n)_{n\in\mathbb{N}} = (1, 4, 9, 16, \dots)$.

Complex sequences:

d) $d_n = \frac{1}{n} + in^2, \quad (d_n)_{n\in\mathbb{N}} = (1 + i, \frac{1}{2} + 4i, \frac{1}{3} + 9i, \frac{1}{4} + 16i, \dots)$.

e) $e_n = i^n, \quad (e_n)_{n\in\mathbb{N}} = (i, -1, -i, 1, i, \dots)$.

Without further ado, we introduce the notion of convergence that is one of the most important definitions of the entire lecture course since a lot of what is to come will be based on it.

Definition 2.3. Let $K = \mathbb{R}$ or $K = \mathbb{C}$, let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K and let $a \in K$. We say that $(a_n)_{n\in\mathbb{N}}$ *converges to a* if

$$\forall \varepsilon > 0 \ \exists n_0 \in \mathbb{N} : \ |a_n - a| < \varepsilon \ \forall n \geq n_0.$$

In this case, a is called the *limit* of $(a_n)_{n\in\mathbb{N}}$ and we denote it by

$$\lim_{n\to\infty} a_n = a.$$

A sequence in K is called *convergent* if it converges to some limit in K. A sequence in K that is not convergent is called *divergent*.

Theorem 2.4. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$. The limit of a convergenct sequence in K is unique.*

Proof. Let $a, b \in K$ and let $(a_n)_{n\in\mathbb{N}}$ converge both to a and to b. Let $\varepsilon > 0$ be arbitrary. By assumption there exist $n_1, n_2 \in \mathbb{N}$, such that

$$\begin{aligned} |a_n - a| &< \varepsilon \quad \forall n \geq n_1, \\ |a_n - b| &< \varepsilon \quad \forall n \geq n_2. \end{aligned}$$

If we put $n_0 := \max\{n_1, n_2\}$, then we obtain for each $n \geq n_0$ using the triangle inequality that

$$\begin{aligned} |a - b| &= |a - a_n + a_n - b| \\ &\leq |a - a_n| + |a_n - b| = |a_n - a| + |a_n - b| < \varepsilon + \varepsilon = 2\varepsilon. \end{aligned}$$

So we have shown that $|a - b| < 2\varepsilon$ for all $\varepsilon > 0$. If we had $a \neq b$, then $|a - b| > 0$, so choosing $\varepsilon = \frac{|a-b|}{2}$ would yield $|a - b| < |a - b|$, which is a contradiction. Hence, $a = b$ which shows the claim. □

Example 2.5. (1) Let $c \in \mathbb{C}$. Then the constant sequence $a_n = c$ for all $n \in \mathbb{N}$, converges against c. This is obvious.

(2) The real sequence $(\frac{1}{n})_{n\in\mathbb{N}}$ converges to 0:

Let $\varepsilon > 0$ be arbitrary. By Corollary 1.53 there exists $n_0 \in \mathbb{N}$ with $\frac{1}{n_0} < \varepsilon$. Since for $n \geq n_0$ it holds that $\frac{1}{n} \leq \frac{1}{n_0}$, we derive that

$$\left|\frac{1}{n} - 0\right| = \frac{1}{n} \leq \frac{1}{n_0} < \varepsilon \qquad \forall n \geq n_0.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows the claim.

(3) The real sequence $((-1)^n)_{n\in\mathbb{N}}$ does not converge:

Assume the sequence converges against $a \in \mathbb{R}$. Since the sequence is given as $(-1, 1, -1, 1, \dots)$, this would particularly imply that for each $\varepsilon > 0$:

$$|1 - a| < \varepsilon, \qquad |-1 - a| = |1 + a| < \varepsilon.$$

Hence, using the triangle inequality:

$$2\varepsilon > |1 - a| + |1 + a| \geq |1 - a + 1 + a| = 2,$$

which is a contradiction if we choose $0 < \varepsilon < 1$. Thus, the sequence diverges.

(4) The real sequence $(\frac{n}{n+1})_{n\in\mathbb{N}}$ converges to 1:

We compute for every $n \in \mathbb{N}$ that

$$\left|\frac{n}{n+1} - 1\right| = \left|\frac{n-(n+1)}{n+1}\right| = \frac{1}{n+1} < \frac{1}{n}.$$

Given $\varepsilon > 0$ let $n_0 \in \mathbb{N}$ with $\frac{1}{n_0} < \varepsilon$. It then follows that

$$\left|\frac{n}{n+1} - 1\right| < \frac{1}{n} \leq \frac{1}{n_0} < \varepsilon \quad \forall n \geq n_0.$$

This shows the claim.

Another important example that takes a little more effort to prove is the following.

Theorem 2.6. *Let $q \in \mathbb{C}$ with $|q| < 1$. Then the sequence $a_n = q^n$, $n \in \mathbb{N}$, converges against 0.*

Proof. We need to show that for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ with $|q^n| < \varepsilon$ for all $n \geq n_0$.
If $q = 0$, then the sequence is constant and the claim immediately follows, so we might assume that $q \neq 0$. Since $|q| < 1$, it holds that $\frac{1}{|q|} > 1$, i.e. there exists $h \in \mathbb{R}$ with $h > 0$, such that $\frac{1}{|q|} = 1 + h$. Then for each $n \in \mathbb{N}$ we derive that

$$|q^n| = |q|^n = \frac{1}{(\frac{1}{|q|})^n} = \frac{1}{(1+h)^n} \overset{(*)}{\leq} \frac{1}{1+nh} < \frac{1}{nh},$$

where the inequality $(*)$ follows from Bernoulli's inequality (Theorem 1.63).
Let $\varepsilon > 0$ be arbitrary. By Corollary 1.53 there exists $n_0 \in \mathbb{N}$, such that

$$\frac{1}{n_0} < h \cdot \varepsilon \quad \Leftrightarrow \quad \frac{1}{n_0 h} < \varepsilon.$$

Then for all $n \geq n_0$ it follows from the above inequalities that

$$|q^n - 0| = |q^n| < \frac{1}{nh} \leq \frac{1}{n_0 h} < \varepsilon,$$

which proves the statement. □

We next investigate how the notion of convergence behaves with respect to addition and multiplication in $\mathbb{R}$ or $\mathbb{C}$.

Theorem 2.7. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be convergent sequences in K with $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$.*

a) $(a_n + b_n)_{n\in\mathbb{N}}$ converges in K with $\lim_{n\to\infty}(a_n + b_n) = a + b$.

b) For each $c \in K$, the sequence $(c \cdot a_n)_{n\in\mathbb{N}}$ converges in K with $\lim_{n\to\infty}(ca_n) = ca$.

Proof. a) Let $\varepsilon > 0$ be arbitrary. By assumption, there exist $n_1, n_2 \in \mathbb{N}$ with $|a_n - a| < \frac{\varepsilon}{2}$ for all $n \geq n_1$ and $|b_n - b| < \frac{\varepsilon}{2}$ for all $n \geq n_2$. Putting $n_0 := \max\{n_1, n_2\}$, we obtain

$$|(a_n + b_n) - (a + b)| = |a_n - a + b_n - b| \leq |a_n - a| + |b_n - b| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad \forall n \geq n_0.$$

Thus, $(a_n + b_n)_{n\in\mathbb{N}}$ converges to $a + b$.

b) If $c = 0$, then the sequence is constantly 0 and the claim is obvious, so assume that $c \neq 0$. Let again $\varepsilon > 0$ be arbitrary. By assumption, there exists $n_3 \in \mathbb{N}$ with $|a_n - a| < \frac{\varepsilon}{|c|}$ for all $n \geq n_3$. It then follows that

$$|ca_n - ca| = |c(a_n - a)| = |c| \cdot |a_n - a| < |c| \cdot \frac{\varepsilon}{|c|} = \varepsilon \quad \forall n \geq n_3.$$

Thus, $(ca_n)_{n\in\mathbb{N}}$ converges to ca.

□

Remark 2.8. It particularly follows from Theorem 2.7 that a sequence $(a_n)_{n\in\mathbb{N}}$ converges to $a \in K$ if and only if $(a_n - a)_{n\in\mathbb{N}}$ converges to 0. (To see "⇒", apply part a) with $b_n = a$ for all $n \in \mathbb{N}$ and part b) with $c = -1$, the other direction follows analogously.)

Before investigating how products of sequences behave, we introduce another important property of sequences.

Definition 2.9. Let $K = \mathbb{R}$ or $K = \mathbb{C}$. A sequence $(a_n)_{n\in\mathbb{N}}$ in K is called *bounded* if there exists $c \geq 0$ with

$$|a_n| \leq c \qquad \forall n \in \mathbb{N},$$

i.e. if $\{|a_n| \mid n \in \mathbb{N}\} \subset \mathbb{R}$ is bounded from above. A sequence which is not bounded is called *unbounded*.

Theorem 2.10. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$. Every convergent sequence in K is bounded.*

Proof. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K with $\lim_{n\to\infty} a_n = a \in K$. In particular, there exists $n_0 \in \mathbb{N}$ such that

$$|a_n - a| < 1 \qquad \forall n \geq n_0.$$

Consequently, by the triangle inequality,

$$|a_n| = |a_n - a + a| \leq |a_n - a| + |a| < |a| + 1 \quad \forall n \geq n_0.$$

Hence, if we put

$$c := \max\{|a_1|, |a_2|, \ldots, |a_{n_0} - 1|, |a| + 1\},$$

it follows that $|a_n| \leq c$ for all $n \in \mathbb{N}$, such that $(a_n)_{n\in\mathbb{N}}$ is bounded. □

Example 2.11. Consider the sequence $(q^n)_{n\in\mathbb{N}}$ for $q \in \mathbb{C}$ with $|q| > 1$. Then for each $n \in \mathbb{N}$,

$$|q^n| = |q|^n = (1 + (|q| - 1))^n \geq 1 + n(|q| - 1)$$

by Bernoulli's inequality. This implies that $(q^n)_{n\in\mathbb{N}}$ is unbounded, hence divergent by Theorem 2.10.

Note that the converse statement of Theorem 2.10 is false. The sequence $((-1)^n)_{n\in\mathbb{N}}$ is bounded, but not convergent.

Theorem 2.12. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be sequences in K.*

a) If $(a_n)_{n\in\mathbb{N}}$ is bounded and $(b_n)_{n\in\mathbb{N}}$ converges to 0, then $(a_n \cdot b_n)_{n\in\mathbb{N}}$ converges to 0.

b) If $(a_n)_{n\in\mathbb{N}}$ converges to a and $(b_n)_{n\in\mathbb{N}}$ converges to b, then $(a_n \cdot b_n)_{n\in\mathbb{N}}$ converges to $a \cdot b$.

Proof. a) Since $(a_n)_{n\in\mathbb{N}}$ is bounded, there exists $c \in \mathbb{R}$ with $|a_n| \leq c$ for all $n \in \mathbb{N}$.

Let $\varepsilon > 0$. Since $(b_n)_{n\in\mathbb{N}}$ converges to 0, there exists $n_0 \in \mathbb{N}$, such that $|b_n| < \frac{\varepsilon}{c}$ for all $n \geq n_0$. Consequently,

$$|a_n \cdot b_n| = |a_n| \cdot |b_n| < c \cdot \frac{\varepsilon}{c} = \varepsilon \qquad \forall n \geq n_0.$$

Thus, $(a_n \cdot b_n)_{n\in\mathbb{N}}$ converges to 0.

b) Let $c_n := b_n - b$ for all $n \in \mathbb{N}$. By Remark 2.8, it holds that $\lim_{n\to\infty} c_n = 0$. Since $(a_n)_{n\in\mathbb{N}}$ is convergent, hence bounded by Theorem 2.10, it follows from part a) that $(a_n \cdot c_n)_{n\in\mathbb{N}}$ converges to 0. Using this observation, we compute that

$$\begin{aligned}\lim_{n\to\infty}(a_n \cdot b_n) &= \lim_{n\to\infty}(a_n \cdot (b_n - b) + a_n \cdot b) \\ &\overset{\text{Theorem 2.10.a)}}{=} \lim_{n\to\infty}(a_n \cdot c_n) + \lim_{n\to\infty}(a_n \cdot b) \\ &\overset{\text{Theorem 2.10.b)}}{=} 0 + (\lim_{n\to\infty} a_n) \cdot b \\ &= a \cdot b.\end{aligned}$$

This shows the claim.

□

Remark 2.13. The assumption that $(a_n)_{n\in\mathbb{N}}$ is bounded is indeed necessary in Theorem 2.12.a) as the following example shows:
Let $a_n = n^2$ and $b_n = \frac{1}{n}$ for all $n \in \mathbb{N}$. Then $(a_n)_{n\in\mathbb{N}}$ is unbounded, while $\lim_{n\to\infty} b_n = 0$. Then $a_n \cdot b_n = n$, such that $(a_n \cdot b_n)_{n\in\mathbb{N}}$ is unbounded, thus divergent by Theorem 2.10.

Theorem 2.14. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K with $a_n \neq 0$ for every $n \in \mathbb{N}$. If $\lim_{n\to\infty} a_n = a$, where $a \in K$ with $a \neq 0$, then $(\frac{1}{a_n})_{n\in\mathbb{N}}$ is convergent with*

$$\lim_{n\to\infty} \frac{1}{a_n} = \frac{1}{a}.$$

Proof. By Remark 2.8, it suffices to show that $(\frac{1}{a_n} - \frac{1}{a})_{n\in\mathbb{N}}$ converges to 0. For each $n \in \mathbb{N}$ we compute that

$$\frac{1}{a_n} - \frac{1}{a} = \frac{a - a_n}{a_n a} = (a - a_n)\frac{1}{a_n a}.$$

Since $\lim_{n\to\infty} a_n = a$, it holds by Remark 2.8 and Theorem 2.7.b) that $\lim_{n\to\infty}(a - a_n) = -\lim_{n\to\infty}(a_n - a) = 0$. Thus, the claim will follow from Theorem 2.12.a) if we can show that $(\frac{1}{a_n a})_{n\in\mathbb{N}}$ is bounded. Since $(a_n - a)_{n\in\mathbb{N}}$ converges to 0, there particularly exists $n_0 \in \mathbb{N}$ such that

$$|a_n - a| < \frac{|a|}{2} \qquad \forall n \geq n_0. \tag{2.1}$$

By the triangle inequality, $|a| \leq |a_n - a| + |a_n|$ and thus

$$|a_n| \geq |a| - |a_n - a| \qquad \forall n \in \mathbb{N}. \tag{2.2}$$

Combining (2.1) and (2.2) yields

$$|a_n| \geq |a| - |a_n - a| > |a| - \frac{|a|}{2} = \frac{|a|}{2} \qquad \forall n \geq n_0.$$

Consequently,

$$\left|\frac{1}{a_n a}\right| = \frac{1}{|a_n| \cdot |a|} \leq \frac{2}{|a|^2} \qquad \forall n \geq n_0.$$

Thus, $|\frac{1}{a_n a}| \leq C$ for all $n \in \mathbb{N}$, where $C := \max\{\frac{1}{|a_1 a|}, \ldots, \frac{1}{|a_{n_0-1} a|}, \frac{2}{|a|^2}\}$. Hence, $(\frac{1}{a_n a})_{n\in\mathbb{N}}$ is bounded. □

Remark 2.15. Combining Theorems 2.12 and 2.14, we derive that if $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are convergent sequences in K, where $K = \mathbb{R}$ or $K = \mathbb{C}$, for which $b_n \neq 0$ for all $n \in \mathbb{N}$ and $\lim_{n\to\mathbb{N}} b_n \neq 0$, then

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n}.$$

Example 2.16. Let

$$a_n = \frac{3n^2 + 5n}{n^2 + 1}, \quad n \in \mathbb{N}.$$

Here, both the sequence in the numerator and the one in the denumerator are unbounded. But reducing fractions, we obtain that $a_n = \frac{3+\frac{5}{n}}{1+\frac{1}{n^2}}$ for each $n \in \mathbb{N}$. Here, we see that the sequences in numerator and denominator converge and we obtain

$$\lim_{n\to\infty} \frac{3n^2 + 5n}{n^2 + 1} = \lim_{n\to\infty} \frac{3 + \frac{5}{n}}{1 + \frac{1}{n^2}} = \frac{3 + \lim_{n\to\infty} \frac{5}{n}}{1 + \lim_{n\to\infty} \frac{1}{n^2}} = \frac{3}{1} = 3.$$

In general, when considering the convergence properties of a sequence of fractions for which both numerator and denominator are sums of powers of n, it is the recommended strategy to *reduce the fraction by the highest power of n that occurs in the denominator*.

Definition 2.17. Let X be a set and $(a_n)_{n\in\mathbb{N}}$ be a sequence in X. A sequence $(b_k)_{k\in\mathbb{N}}$ in X is called *a subsequence of* $(a_n)_{n\in\mathbb{N}}$ if there exists a map $\varphi : \mathbb{N} \to \mathbb{N}$ with

$$\varphi(1) < \varphi(2) < \cdots < \varphi(n) < \varphi(n+1) < \ldots,$$

such that

$$b_k = a_{\varphi(k)} \qquad \forall k \in \mathbb{N}.$$

In this case, one also puts $n_k = \varphi(k)$ for every $k \in \mathbb{N}$ and denotes the subsequence by $(a_{n_k})_{k\in\mathbb{N}}$.

Example 2.18. (1) The sequence $(\frac{1}{n^2})_{n\in\mathbb{N}}$ is a subsequence of $(\frac{1}{n})_{n\in\mathbb{N}}$. Here, $\varphi : \mathbb{N} \to \mathbb{N}$, $\varphi(n) = n^2$.

(2) The constant sequence $(1)_{n\in\mathbb{N}}$ is a subsequence of $((-1)^n)_{n\in\mathbb{N}}$. Here, $\varphi(n) = 2n$.

(3) The sequence $((-1)^n)_{n\in\mathbb{N}}$ is a subsequence of $(i^n)_{n\in\mathbb{N}}$. Here, $\varphi(n) = 2n + 2$.

Theorem 2.19. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K which converges to $a \in K$. Then every subsequence of $(a_n)_{n\in\mathbb{N}}$ converges to a.*

Proof. Let $(a_{n_k})_{k\in\mathbb{N}}$ be a subsequence of $(a_n)_{n\in\mathbb{N}}$. Since by assumption

$$n_1 < n_2 < \cdots < n_k < n_{k+1} < \ldots,$$

it holds that $n_k \geq k$ for each $k \in \mathbb{N}$.

Let $\varepsilon > 0$. By assumption there exists $k_0 \in \mathbb{N}$, such that $|a_n - a| < \varepsilon$ for all $n \geq k_0$. Then it particularly follows for all $k \geq k_0$ from $n_k \geq k \geq k_0$ that

$$|a_{n_k} - a| < \varepsilon \qquad \forall k \geq k_0.$$

Since ε was arbitrary, it follows that $(a_{n_k})_{k\in\mathbb{N}}$ converges to a. □

For any complex sequence $(z_n)_{n\in\mathbb{N}}$ we have the two associated real sequences $(\mathrm{Re}(z_n))_{n\in\mathbb{N}}$ and $(\mathrm{Im}(z_n))_{n\in\mathbb{N}}$. The following result shows that convergence of $(z_n)_{n\in\mathbb{N}}$ is equivalent to the convergence of both of these sequences.

Theorem 2.20. *Let $(c_n)_{n\in\mathbb{N}}$ be a complex sequence and let $c_n = a_n + ib_n$, where $a_n, b_n \in \mathbb{R}$ for all $n \in \mathbb{N}$. Let $c = a + ib \in \mathbb{C}$. Then the following holds: $(c_n)_{n\in\mathbb{N}}$ converges to c if and only if $(a_n)_{n\in\mathbb{N}}$ converges to a and $(b_n)_{n\in\mathbb{N}}$ converges to b.*

Proof. "⇒": Let $(c_n)_{n\in\mathbb{N}}$ converge against c, so for every $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ with

$$|c_n - c| < \varepsilon \quad \forall n \geq n_0.$$

Then it follows from Theorem 1.73.c) that

$$|a_n - a| = |\mathrm{Re}(c_n - c)| \leq |c_n - c| < \varepsilon, \quad |b_n - b| = |\mathrm{Im}(c_n - c)| \leq |c_n - c| < \varepsilon \qquad \forall n \geq n_0.$$

The convergence of $(a_n)_{n\to\infty}$ to a and of $(b_n)_{n\in\mathbb{N}}$ to b immediately follow.
"⇐": We first observe that for every $z \in \mathbb{C}$ the triangle inequality implies: In particular, for all $n \in \mathbb{N}$ it holds that

$$|c_n - c| \leq |a_n - a| + |b_n - b|. \tag{2.3}$$

Let $\varepsilon > 0$. By assumption, there exist $n_1, n_2 \in \mathbb{N}$ with $|a_n - a| < \frac{\varepsilon}{2}$ for all $n \geq n_1$ and $|b_n - b| < \frac{\varepsilon}{2}$ for all $n \geq n_2$. Putting $n_0 := \max\{n_1, n_2\}$ we thus obtain from (2.3) that

$$|c_n - c| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \forall n \geq n_0.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows the convergence of $(c_n)_{n\in\mathbb{N}}$ to c. □

Remark 2.21. Note that in the case of convergence, the previous theorem says that

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} a_n + i \lim_{n\to\infty} b_n.$$

2.2 Convergence of real sequences

In this section and the next one, we want to focus on sequences in $\mathbb{R}$. The ordering properties of $\mathbb{R}$ can be used in various ways to establish convergence of real sequences as we shall discuss in this section. Before computing an important limit we discuss a property of roots that we will need in our considerations.

Theorem 2.22. *Let $x, y \in \mathbb{R}$ with $0 \leq x < y$ and let $n \in \mathbb{N}$. Then $\sqrt[n]{x} < \sqrt[n]{y}$.*

Proof. Assume that there were $n \in \mathbb{N}$ and $0 < x < y$ with $\sqrt[n]{x} \geq \sqrt[n]{y}$. Since $\sqrt[n]{x}$ and $\sqrt[n]{y}$ are non-negative, it then follows from the properties of ordered fields that

$$(\sqrt[n]{x})^n \geq (\sqrt[n]{y})^n \quad \Leftrightarrow \quad x \geq y.$$

This is a contradiction, which proves the statement. □

Example 2.23. For all $p, q \in \mathbb{N}$, it holds that $\lim_{n\to\infty} \frac{1}{\sqrt[q]{n^p}} = 0$. (Here, $\sqrt[1]{x} := x$ for all $x \geq 0$.)
We first note that whenever $x < y$, then $\sqrt[q]{x} < \sqrt[q]{y}$ for all $q \in \mathbb{N}$ as we shall see in Theorem 2.22 below.
Let $\varepsilon > 0$. There exists $n_0 \in \mathbb{N}$ such that $\frac{1}{n_0} < \varepsilon^{\frac{q}{p}}$, hence

$$\frac{1}{\sqrt[q]{n_0^p}} = \left(\frac{1}{n_0}\right)^{\frac{p}{q}} < \varepsilon$$

by Theorem 2.22. By the same theorem it follows that

$$\left|\frac{1}{\sqrt[q]{n^p}}\right| = \frac{1}{\sqrt[q]{n^p}} \leq \frac{1}{\sqrt[q]{n_0^p}} < \varepsilon \qquad \forall n \geq n_0.$$

This shows that $\lim_{n\to\infty} \frac{1}{\sqrt[q]{n^p}} = 0$.

We next compute two more interesting limits involving roots.

Theorem 2.24. *a) Let $a \in \mathbb{R}$ with $a > 0$. Then* $\lim\limits_{n\to\infty} \sqrt[n]{a} = 1$.

b) $\lim\limits_{n\to\infty} \sqrt[n]{n} = 1$.

Proof. a) We first consider the case of $a \geq 1$. Then $\sqrt[n]{a} \geq 1$ for each $n \in \mathbb{N}$. Put $x_n := \sqrt[n]{a} - 1$ for all $n \in \mathbb{N}$. Applying Bernoulli's inequality, we obtain $a = (1 + x_n)^n \geq 1 + nx_n$, hence

$$0 \leq x_n \leq \frac{a-1}{n} < \frac{a}{n} \qquad \forall n \in \mathbb{N}.$$

Given $\varepsilon > 0$ let $n_0 \in \mathbb{N}$ be so big that $\frac{1}{n_0} < \frac{\varepsilon}{a}$. Then for all $n \geq n_0$ we obtain

$$|\sqrt[n]{a} - 1| = x_n < \frac{a}{n} \leq \frac{a}{n_0} < \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows the convergence of $(\sqrt[n]{a})_{n\in\mathbb{N}}$ against 1.

If $a < 1$, it holds that $a^{-1} > 1$ and one easily derives that $\sqrt[n]{a^{-1}} = \frac{1}{\sqrt[n]{a}}$ for each $n \in \mathbb{N}$. Hence, with the help of Theorem 2.14, we obtain

$$\lim_{n\to\infty} \sqrt[n]{a} = \lim_{n\to\infty} \frac{1}{\frac{1}{\sqrt[n]{a}}} = \frac{1}{\lim_{n\to\infty} \sqrt[n]{a^{-1}}} = 1.$$

b) Let $x_n := \sqrt[n]{n} - 1 \geq 0$ for each $n \in \mathbb{N}$. Using the binomial theorem (Theorem 1.64) we obtain for each $n \in \mathbb{N}$:

$$\begin{aligned} n = (1 + x_n)^n &= \sum_{k=0}^{n} \binom{n}{k} 1^k x_n^{n-k} \\ &= \binom{n}{n} + \binom{n}{n-1} x_n + \binom{n}{n-2} x_n^2 + \sum_{k=3}^{n} \binom{n}{k} 1^k x_n^{n-k}. \end{aligned}$$

We compute that $\binom{n}{n} = \frac{n!}{n!0!} = 1$, $\binom{n}{n-1} = \frac{n!}{(n-1)!\cdot 1!} = n$, $\binom{n}{n-2} = \frac{n!}{(n-2)!2!} = \frac{n(n-1)}{2}$, and derive:

$$\begin{aligned} n &= 1 + nx_n + \frac{n(n-1)}{2}x_n^2 + \sum_{k=0}^{n-3}\binom{n}{k}x_n^{n-k} \\ &\geq 1 + \frac{n(n-1)}{2}x_n^2. \\ \Rightarrow \quad n-1 &\geq \frac{n(n-1)}{2}x_n^2. \end{aligned}$$

Thus, for each $n \geq 2$ we obtain

$$x_n^2 \leq \frac{2}{n} \quad \Rightarrow \quad x_n \leq \sqrt{\frac{2}{n}},$$

where we used $x_n \geq 0$ in the final implication.

Let $\varepsilon > 0$ and let $n_0 \in \mathbb{N}$ be so big that $\frac{1}{n_0} < \frac{\varepsilon^2}{2}$. Then for each $n \geq n_0$ we obtain using Theorem 2.22

$$|\sqrt[n]{n} - 1| = x_n \leq \sqrt{\frac{2}{n}} \leq \sqrt{\frac{2}{n_0}} < \sqrt{\varepsilon^2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows the claim.

□

Definition 2.25. Let X be a set and $(a_n)_{n\in\mathbb{N}}$ be a sequence in X. We say that the elements of $(a_n)_{n\in\mathbb{N}}$ has a property *for almost every* $n \in \mathbb{N}$, if all but finitely many of the a_n have this property, or equivalently, if there exists an $n_0 \in \mathbb{N}$, such that a_n has the property for every $n \geq n_0$.

Theorem 2.26. *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be convergent real sequences. If $a_n \leq b_n$ for almost every $n \in \mathbb{N}$, then*

$$\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n.$$

Proof. Put $a := \lim_{n\to\infty} a_n$ and $b := \lim_{n\to\infty} b_n$. Let $n_1 \in \mathbb{N}$, such that $a_n \leq b_n$ for all $n \geq n_1$. Given ε, there exist $n_2, n_3 \in \mathbb{N}$ with $|a_n - a| < \varepsilon$ for all $n \geq n_2$ and $|b_n - b| < \varepsilon$ for all $n \geq n_3$. Putting $n_0 := \max\{n_1, n_2, n_3\}$ we derive that

$$a - \varepsilon < a_n \leq b_n < b + \varepsilon \qquad \forall n \geq n_0.$$

Thus, $a - b < 2\varepsilon$ for all $\varepsilon > 0$. Similar to the proof of Theorem 2.4, this implies $a - b \leq 0$, or $a \leq b$. □

Corollary 2.27. *Let $(a_n)_{n\in\mathbb{N}}$ be a convergent real sequence and let $c \in \mathbb{R}$.*

a) If $a_n \leq c$ for almost every $n \in \mathbb{N}$, then $\lim_{n\to\infty} a_n \leq c$.

b) If $a_n \geq c$ for almost every $n \in \mathbb{N}$, then $\lim_{n\to\infty} a_n \geq c$.

Proof. This follows from Theorem 2.26 by letting one of the two sequences be the constant sequence $(c)_{n\in\mathbb{N}}$. □

The following theorem is a very useful tool to show the convergence of sequences that are "caught between" two convergent sequences.

Theorem 2.28 (Squeeze Theorem[1]). *Let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ be real sequences with*

$$a_n \leq b_n \leq c_n$$

for almost every $n \in \mathbb{N}$. If $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ are convergent with $\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n =: a$, then $(b_n)_{n\in\mathbb{N}}$ converges to a as well.

Proof. Let $\varepsilon > 0$. In analogy with the proof of Theorem 2.26 one shows that there exists $n_0 \in \mathbb{N}$ with

$$a - \varepsilon < a_n \leq b_n \leq c_n < a + \varepsilon \qquad \forall n \geq n_0.$$

Hence, $b_n - a < \varepsilon$ and $a - b_n < \varepsilon$ for all $n \geq n_0$, in other words

$$|b_n - a| < \varepsilon \qquad \forall n \geq n_0,$$

which shows that $\lim_{n\to\infty} b_n = a$. □

We want to formulate a very common special case of Theorem 2.28 as a corollary of its own.

Corollary 2.29. *Let $(x_n)_{n\in\mathbb{N}}$ be a real sequence with $x_n \geq 0$ for all $n \in \mathbb{N}$. If there exists a real sequence $(c_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} c_n = 0$ and $x_n \leq c_n$ for almost every $n \in \mathbb{N}$, then $\lim_{n\to\infty} x_n = 0$.*

Proof. Apply Theorem 2.28 with $a_n = 0$ and $x_n = b_n$ for all $n \in \mathbb{N}$. □

Example 2.30. (1) $\lim\limits_{n\to\infty} (\sqrt{n+1} - \sqrt{n}) = 0$.

To show this, we use a little computational trick. For each $n \in \mathbb{N}$ we compute:

$$\sqrt{n+1} - \sqrt{n} = \frac{(\sqrt{n+1} - \sqrt{n})(\sqrt{n+1} + \sqrt{n})}{\sqrt{n+1} + \sqrt{n}} = \frac{n+1-n}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}}.$$

Hence,

$$0 \leq \sqrt{n+1} - \sqrt{n} = \frac{1}{\sqrt{n+1} + \sqrt{n}} \leq \frac{1}{\sqrt{n}} \qquad \forall n \in \mathbb{N}.$$

By Example 2.23.(2) and Corollary 2.29, this shows the claim.

(2) $\lim\limits_{n\to\infty} (\sqrt{n+1} - \sqrt{n})\sqrt{n} = \dfrac{1}{2}$.

Here, the same trick as in (1) shows that

$$(\sqrt{n+1} - \sqrt{n})\sqrt{n} = \frac{\sqrt{n}}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\frac{\sqrt{n+1}}{\sqrt{n}} + 1} = \frac{1}{\sqrt{1 + \frac{1}{n}} + 1}$$

for every $n \in \mathbb{N}$. It further follows from Exercise ... of Sheet 3 that $\lim_{n\to\infty} \sqrt{1 + \frac{1}{n}} = 1$. Hence, by the computational rules for limits from the previous section,

$$\lim_{n\to\infty} (\sqrt{n+1} - \sqrt{n})\sqrt{n} = \frac{1}{\lim_{n\to\infty} \sqrt{1 + \frac{1}{n}} + 1} = \frac{1}{2}.$$

[1] This theorem has many interesting and slightly different names in different countries: Einschließungssatz (Germany), sandwich theorem (United Kingdom), teorema dei due carabiniere (Italy), théoreme des gendarmes, teorema del emparedado (Spain), ...

(3) For every $z \in \mathbb{C}$ it holds that $\lim_{n\to\infty} \frac{z^n}{n!} = 0$.

For $z = 0$ the statement is obvious, so assume that $z \neq 0$. We then compute for each $n \in \mathbb{N}$ with $a_n = \frac{|z|^n}{n!}$ that

$$\frac{a_{n+1}}{a_n} = \frac{|z|^{n+1} \cdot n!}{(n+1)! \cdot |z|^n} = \frac{|z|}{n+1}.$$

Thus, there exists $n_0 \in \mathbb{N}$ with $\frac{a_{n+1}}{a_n} \leq \frac{1}{2}$ for all $n \geq n_0$. Iterating this inequality, i.e. applying it again and again, we obtain that

$$0 \leq a_n \leq \left(\frac{1}{2}\right)^{n-n_0} a_{n_0} \qquad \forall n \geq n_0.$$

Since $\lim_{n\to\infty} \left(\frac{1}{2}\right)^{n-n_0} a_{n_0} = (\frac{1}{2})^{-n_0} a_{n_0} \cdot \lim_{n\to\infty} (\frac{1}{2})^n = 0$ by Theorem 2.6, it follows from the Squeeze Theorem that

$$\lim_{n\to\infty} a_n = 0,$$

which implies that $\lim_{n\to\infty} \frac{z^n}{n!} = 0$ as well.

We next want to establish a very helpful criterion for the convergence of real sequences. To do so, we first need to introduce some additional terminology.

Definition 2.31. A real sequence $(a_n)_{n\in\mathbb{N}}$ is *bounded from above* if there exists $C \in \mathbb{R}$ with $a_n \leq C$ for all $n \in \mathbb{N}$. $(a_n)_{n\in\mathbb{N}}$ is *bounded from below* if there exists $c \in \mathbb{R}$ with $a_n \geq c$ for all $n \in \mathbb{N}$.
In other words, $(a_n)_{n\in\mathbb{N}}$ is bounded from above (below) if the set $\{a_n \mid n \in \mathbb{N}\} \subset \mathbb{R}$ is bounded from above (below).

Example 2.32. (1) Every bounded real sequence is bounded from above and bounded from below.

(2) $a_n = n$, $n \in \mathbb{N}$, is not bounded from above by Theorem 1.52, but bounded from below, e.g. by 1.

Definition 2.33. Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence.

a) $(a_n)_{n\in\mathbb{N}}$ is *monotonically increasing* if $a_n \leq a_{n+1}$ for all $n \in \mathbb{N}$.

b) $(a_n)_{n\in\mathbb{N}}$ is *monotonically decreasing* if $a_n \geq a_{n+1}$ for all $n \in \mathbb{N}$.

c) $(a_n)_{n\in\mathbb{N}}$ is called *monotonic* if it is monotonically increasing or monotonically decreasing.

Example 2.34. (1) Every constant sequence is monotonically increasing and monotonically decreasing.

(2) $(n)_{n\in\mathbb{N}}$ is monotonically increasing while $(-n)_{n\in\mathbb{N}}$ is monotonically decreasing.

(3) $a_n = (-1)^n$, $n \in \mathbb{N}$, is neither monotonically increasing nor monotonically decreasing, since $a_{2n-1} \leq a_{2n}$, but $a_{2n} \geq a_{2n+1}$ for every $n \in \mathbb{N}$.

Theorem 2.35. *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence.*

a) If $(a_n)_{n\in\mathbb{N}}$ is monotonically increasing and bounded from above, then $(a_n)_{n\in\mathbb{N}}$ converges with

$$\lim_{n\to\infty} a_n = \sup\{a_n \mid n \in \mathbb{N}\}.$$

b) If $(a_n)_{n\in\mathbb{N}}$ is monotonically decreasing and bounded from below, then $(a_n)_{n\in\mathbb{N}}$ converges with

$$\lim_{n\to\infty} a_n = \inf\{a_n \mid n \in \mathbb{N}\}.$$

Proof. a) Since $\{a_n \mid n \in \mathbb{N}\}$ is bounded from above, it has a supremum $S := \sup\{a_n \mid n \in \mathbb{N}\}$ by completeness of $\mathbb{R}$.

Let $\varepsilon > 0$. Since $S - \varepsilon$ is not an upper bound for $\{a_n \mid n \in \mathbb{N}\}$, there exists $n_0 \in \mathbb{N}$ with

$$S - \varepsilon < a_{n_0} \leq S.$$

Since $(a_n)_{n\in\mathbb{N}}$ is monotonically increasing and S is an upper bound, this implies

$$S - \varepsilon < a_n \leq S \qquad \forall n \geq n_0.$$

Consequently, $0 \leq S - a_n < \varepsilon$ for each $n \geq n_0$, such that

$$|a_n - S| = S - a_n < \varepsilon \qquad \forall n \geq n_0.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that $\lim_{n\to\infty} a_n = S$.

b) Consider the sequence $b_n = -a_n$, $n \in \mathbb{N}$. Then $(b_n)_{n\in\mathbb{N}}$ is monotonically increasing and bounded from above, such that by a) it converges with

$$\lim_{n\to\infty} b_n = \sup\{b_n \mid n \in \mathbb{N}\}.$$

But in analogy with the proof of Theorem 1.45 one shows that

$$\sup\{b_n \mid n \in \mathbb{N}\} = -\inf\{a_n \mid n \in \mathbb{N}\}.$$

Together with Theorem 2.7.b) this yields the convergence of $(a_n)_{n\in\mathbb{N}}$ with

$$\lim_{n\to\infty} a_n = -\lim_{n\to\infty} b_n = \inf\{a_n \mid n \in \mathbb{N}\}.$$

□

A class of sequences for which the previous theorem is particularly useful are recursively defined sequences.

Theorem/Definition 2.36. *Let X be a set, $f : X \to X$ be a map and $a \in X$. Then there is a unique sequence in X which satisfies*

$$a_1 = a, \qquad a_{n+1} = f(a_n) \quad \forall n \in \mathbb{N}.$$

A sequence defined by these properties is called a recursive sequence.

Proof. The uniqueness part is obvious. For the existence apply the induction principle to the set $M = \{n \in \mathbb{N} \mid a_n \text{ is well-defined}\}$. □

Example 2.37. (1) Assume that a fixed amount of money, say A€, is put on a bank account with an annual interest rate of p %. The amount of money on the account thus grows by a factor of $(1 + \frac{p}{100})$ per year. Thus, if a_n denotes the amount of money on the account in year n counted in €, we can describe $(a_n)_{n\in\mathbb{N}}$ by

$$a_1 = A, \qquad a_{n+1} = \Big(1 + \frac{p}{100}\Big)a_n \quad \forall n \in \mathbb{N}.$$

One proves by induction that the sequence is given by

$$a_n = \Big(1 + \frac{p}{100}\Big)^{n-1} A.$$

This form is then called the *explicit form* of the recursively defined sequence, i.e. a description of each element in terms of n without referring to other elements of the sequence.

(2) Assume that the bank from (1) is generous and additionally puts 1€ into the account each year. With a_n again denoting the amount of money on the account in year n, this leads to the recursive formula

$$a_1 = A, \qquad a_{n+1} = \Big(1 + \frac{p}{100}\Big) a_n + 1 \quad \forall n \in \mathbb{N}.$$

Here, one proves by induction that the sequence is explicitly given by

$$a_n = \Big(1 + \frac{p}{100}\Big)^{n-1} A + \sum_{k=0}^{n-2} \Big(1 + \frac{p}{100}\Big)^k + 1 \qquad \forall n \geq 2.$$

The following theorem shows a typical application of Theorem 2.35 to a recursive sequence and presents an effcient method to approximate square roots by elementary computations.

Theorem 2.38 (Heron's method of computing square roots)**.** *Let $a > 0$ and consider the recursive sequence defined by*

$$x_1 = 1, \qquad x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \quad \forall n \in \mathbb{N}.$$

Then $(x_n)_{n\in\mathbb{N}}$ is convergent with $\lim_{n\to\infty} x_n = \sqrt{a}$.

Proof. We want to apply Theorem 2.35 to $(x_n)_{n\in\mathbb{N}}$ to show convergence.

(i) $x_n > 0$ for all $n \in \mathbb{N}$.

Proof of (i). (by induction over n)
Base case: By definition, $x_1 = 1 > 0$.

Induction step: Assume that $x_n > 0$ is satisfies for some $n \in \mathbb{N}$. Since $a > 0$, it then holds that $\frac{a}{x_n} > 0$ as well, hence

$$x_{n+1} = \frac{1}{2}\Big(\underbrace{x_n}_{>0} + \underbrace{\frac{a}{x_n}}_{>0} \Big) > 0.$$

This completes the induction step.

(ii) $x_{n+1} \geq \sqrt{a}$ for all $n \in \mathbb{N}$.

Proof of (ii). By (i), it suffices to show that $x_{n+1}^2 \overset{!}{\geq} a$ for all $n \in \mathbb{N}$. Using the recursion formula we compute for all $n \in \mathbb{N}$ that

$$\begin{aligned} x_{n+1}^2 &= \frac{1}{4}\Big(x_n + \frac{a}{x_n}\Big)^2 = \frac{1}{4}\Big(x_n^2 + 2a + \frac{a^2}{x_n^2}\Big) \\ &= a + \frac{1}{4}\Big(x_n^2 - 2a + \frac{a^2}{x_n^2}\Big) = a + \frac{1}{4}\Big(x_n - \frac{a}{x_n}\Big)^2 \geq a. \end{aligned}$$

(iii) $x_n \geq x_{n+1}$for al $n \geq 2$.

Proof of (iii). We compute for each $n \in \mathbb{N}$ that

$$x_n - x_{n+1} = \frac{1}{2}x_n - \frac{1}{2}\frac{a}{x_n} = \frac{1}{2}\Big(x_n - \frac{a}{x_n}\Big) = \frac{1}{2x_n}(x_n^2 - a) \geq 0$$

by (ii).

By (ii) and (iii), $(x_n)_{n\in\mathbb{N}}$ is monotonously decreasing and bounded from below, hence convergent by Theorem 2.35.b). Put $L := \lim_{n\to\infty} x_n$. Since $(x_{n+1})_{n\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$, Theorem 2.19 implies $\lim_{n\to\infty} x_{n+1} = L$. With the rules for limits from the previous section, we derive from the recursion formula that

$$0 = \lim_{n\to\infty}(x_n - x_{n+1}) = \lim_{n\to\infty}\frac{1}{2x_n}(x_n^2 - a) = \frac{L^2 - a}{2L}.$$

Hence L satisfies $L^2 = a$ and from (ii) and Corollary 2.27 it eventually follows that $L = \sqrt{a}$. □

Remark 2.39. The sequences from the previous theorem can indeed be used to approximate square roots by a sequence of rational numbers. Let us compute the first terms of the sequence in the case $a = 2$. Then:

$$\begin{aligned}
x_1 &= 1 \\
x_2 &= \frac{1}{2}(1+2) = \frac{3}{2} = 1,5 \\
x_3 &= \frac{1}{2}\left(\frac{3}{2} + \frac{4}{3}\right) = \frac{17}{12} = 1,4166666... \\
x_4 &= \frac{1}{2}\left(\frac{17}{12} + \frac{24}{17}\right) = \frac{577}{408} = 1,414215686... \qquad \vdots
\end{aligned}$$

Our calculator tells us that $\sqrt{2} = 1,414213562...$ So already the fourth term of the sequence approximates $\sqrt{2}$ quite accurately, it holds that $x_4 - \sqrt{2} \approx 2 \cdot 10^{-6}$. Not bad!

2.3 Divergent sequences and their subsequences

In this section we want to investigate further properties of real or complex sequences that do not require their convergence. An important example is the theorem of Bolzano-Weierstraß which says that any bounded sequence has a convergent subsequence. First of all, we want to formalize the notion of real sequences whose values become arbitrarily big.

Definition 2.40. Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence.

a) We say that $(a_n)_{n\in\mathbb{N}}$ *diverges to* $+\infty$ and write $\lim_{n\to\infty} a_n = +\infty$ if for each $C \in \mathbb{R}$ there exists $n_0 \in \mathbb{N}$ such that $a_n \geq C$ for all $n \geq n_0$.

b) We say that $(a_n)_{n\in\mathbb{N}}$ *diverges to* $-\infty$ and write $\lim_{n\to\infty} a_n = -\infty$ if for every $c \in \mathbb{R}$ there exists $n_0 \in \mathbb{N}$ with $a_n \leq c$ for all $n \geq n_0$.

Remark 2.41. (1) It follows directly from the definition that a sequence that diverges to $+\infty$ or to $-\infty$ is unbounded. However, not every unbounded real sequence diverges to $+\infty$ or to $-\infty$. A counterexample is given by $((-1)^n n)_{n\in\mathbb{N}}$.

(2) It follows directly from the definition that if $(a_n)_{n\to\infty}$ diverges to $+\infty$ (or to $-\infty$), then every subsequence of $(a_n)_{n\in\mathbb{N}}$ diverges to $+\infty$ (or to $-\infty$) as well.

(3) It is easy to see that a real sequence $(a_n)_{n\in\mathbb{N}}$ diverges to $+\infty$ if and only if $(-a_n)_{n\in\mathbb{N}}$ diverges to $-\infty$.

As in the convergence case, we can study inequalities between sequences to derive divergence to $+\infty$ or $-\infty$.

Theorem 2.42. *Let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ be real sequences with $a_n \leq b_n$ for almost every $n \in \mathbb{N}$.*

a) If $\lim_{n\to\infty} a_n = +\infty$, then $\lim_{n\to\infty} b_n = +\infty$.

b) If $\lim_{n\to\infty} b_n = -\infty$, then $\lim_{n\to\infty} a_n = -\infty$.

Proof. This follows immediately from the assumption and Definition 2.40. □

In the case of a convergent sequence with a non-zero limit, we have seen that the sequence of inverses then converges against the inverse of the limit as well. There is an analogue of this result for sequences diverging to $\pm\infty$.

Theorem 2.43. *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with $a_n \neq 0$ for all $n \in \mathbb{N}$.*

a) If $\lim_{n\to\infty} a_n = +\infty$ or $\lim_{n\to\infty} a_n = -\infty$, then $\lim_{n\to\infty} \frac{1}{a_n} = 0$.

b) If $\lim_{n\to\infty} a_n = 0$ and $a_n > 0$ for almost every $n \in \mathbb{N}$, then $\lim_{n\to\infty} \frac{1}{a_n} = +\infty$.

c) If $\lim_{n\to\infty} a_n = 0$ and $a_n < 0$ for almost every $n \in \mathbb{N}$, then $\lim_{n\to\infty} \frac{1}{a_n} = -\infty$.

Proof. a) Assume that $\lim_{n\to\infty} a_n = +\infty$. Let $\varepsilon > 0$. By assumption, there exists $n_0 \in \mathbb{N}$ such that $a_n \geq \frac{1}{\varepsilon}$ for all $n \geq n_0$. Hence,

$$\left|\frac{1}{a_n}\right| = \frac{1}{a_n} < \varepsilon \qquad \forall n \geq n_0.$$

Since $\varepsilon > 0$ was arbitrary, this shows that $\lim_{n\to\infty} \frac{1}{a_n} = 0$.

In the case that $\lim_{n\to\infty} a_n = -\infty$, we can apply the same argument to $(-a_n)_{n\in\mathbb{N}}$ which diverges to $+\infty$ to derive that $\lim_{n\to\infty} \frac{1}{a_n} = -\lim_{n\to\infty} \frac{1}{-a_n} = 0$.

b) Let $C \in \mathbb{R}$ with $C > 0$. By assumption, there exists $n_0 \in \mathbb{N}$, such that $|a_n| < \frac{1}{C}$ for all $n \geq n_0$. Consequently,

$$\frac{1}{a_n} = \frac{1}{|a_n|} > C \qquad \forall n \geq n_0.$$

Since $C > 0$ was chosen arbitrarily, this shows that $\lim_{n\to\infty} \frac{1}{a_n} = +\infty$.

c) This is shown by applying b) to $(-a_n)_{n\in\mathbb{N}}$ and using the observation from Remark 2.41.(2). □

Example 2.44. Combining Theorem 2.43.b) with Exercise 4.a) from Sheet 2 shows that

$$\lim_{n\to\infty} \frac{2^n}{n} = +\infty.$$

To prepare an important result about bounded sequences we introduce certain very common subsets of $\mathbb{R}$ that we will use very frequently throughout the course.

Definition 2.45. For $a, b \in \mathbb{R}$ with $a \leq b$ we define

$$\begin{aligned}
[a,b] &:= \{x \in \mathbb{R} \mid a \leq x \leq b\},\\
(a,b) &:= \{x \in \mathbb{R} \mid a < x < b\},\\
[a,b) &:= \{x \in \mathbb{R} \mid a \leq x < b\},\\
(a,b] &:= \{x \in \mathbb{R} \mid a < x \leq b\},\\
(a,+\infty) &:= \{x \in \mathbb{R} \mid a < x\},\\
[a,+\infty) &:= \{x \in \mathbb{R} \mid a \leq x\},\\
(-\infty,b) &:= \{x \in \mathbb{R} \mid x < b\},\\
(-\infty,b] &:= \{x \in \mathbb{R} \mid x \leq b\}.
\end{aligned}$$

A subset of $\mathbb{R}$ of one of these forms is called an *interval*. Intervals of type (a,b), $(a,+\infty)$ or $(-\infty,b)$ are called *open intervals* while intervals of type $[a,b]$, $[a,+\infty)$ or $(-\infty,b]$ are called *closed intervals*. Intervals of type $[a,b)$ or $(a,b]$ are called *half-closed* or *half-open*[2] intervals.
Intervals of type $[a,b]$, (a,b), $[a,b)$ or $(a,b]$ are called *bounded* intervals, those of type $(a,+\infty)$, $[a,+\infty)$, $(-\infty,b)$ or $(-\infty,b]$ are called *unbounded*.

Remark 2.46. A real sequence $(a_n)_{n\in\mathbb{N}}$ is bounded if and only if there are $a, b \in \mathbb{R}$ such that $x_n \in [a,b]$ for all $n \in \mathbb{N}$. Hence, a real sequence is bounded if and only if it is bounded from above and bounded from below.

The following result about intervals that are nested into one another is derived applying methods from the previous section to the sequences of endpoints of intervals. While it might not look particularly interesting at first sight, it has important applications in analysis of which we will see two.

Theorem 2.47 (Nested interval principle[3])**.** *Let $(I_n)_{n\in\mathbb{N}}$ be sequence of closed bounded intervals, where $I_n = [a_n, b_n]$ for all $n \in \mathbb{N}$, which satisfies*

(i) $I_{n+1} \subset I_n$ *for all* $n \in \mathbb{N}$,

(ii) $\lim\limits_{n\to\infty}(b_n - a_n) = 0$.

Then there exists a unique $c \in \mathbb{R}$ with $c \in I_n$ for all $n \in \mathbb{N}$, i.e. $\bigcap_{n\in\mathbb{N}} I_n = \{c\}$.

Proof. We will show the existence and uniqueness of c separately.
Uniqueness: Let $c_1, c_2 \in \bigcap_{n\in\mathbb{N}} I_n$ with $c_1 \leq c_2$. Then by definition of the intersection, it will hold that

$$a_n \leq c_1 \leq c_2 \leq b_n \qquad \forall n \in \mathbb{N}.$$

This implies that

$$0 \leq c_2 - c_1 \leq b_n - a_n \qquad \forall n \in \mathbb{N}.$$

Thus, it follows from assumption (ii) and Corollary 2.29 that $c_2 - c_1 = 0$, i.e. $c_1 = c_2$.

Existence: By (i), $a_n \leq a_{n+1}$ and $b_n \geq b_{n+1}$ for all $n \in \mathbb{N}$, so $(a_n)_{n\in\mathbb{N}}$ is monotonically increasing while $(b_n)_{n\in\mathbb{N}}$ is monotonically decreasing. Moreover $a_n \leq b_1$ and $b_n \geq a_1$ for all $n \in \mathbb{N}$, hence

[2] Some people see a glass as half full, some see it as half empty...
[3] This result's German name is a typical example for a long German composite word: *Intervallschachtelungsprinzip*.

$(a_n)_{n\in\mathbb{N}}$ is bounded from above while $(b_n)_{n\in\mathbb{N}}$ is bounded from below. Thus, by Theorem 2.35 both $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are convergent. By (ii) and Theorem 2.7,

$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n = \lim_{n\to\infty}(b_n - a_n) = 0,$$

hence $c := \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$. By Theorem 2.35 it further holds that

$$c = \sup\{a_n \,|n \in \mathbb{N}\} = \inf\{b_n \mid n \in \mathbb{N}\} \quad \Rightarrow \quad a_n \leq c \leq b_n \ \forall n \in \mathbb{N} \quad \Rightarrow \quad c \in \bigcap_{n\in\mathbb{N}} I_n.$$

□

As discussed in the previous sections, every convergent sequence is bounded, while not every bounded sequence needs to converge. The following theorem shows that at least there is a convergence property of subsequences of bounded sequences.

Theorem 2.48 (Bolzano-Weierstraß)**.** *Every bounded real sequence has a convergent subsequence.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a bounded real sequence and let $[a_1, b_1]$ be a closed bounded interval with $x_n \in [a_1, b_1]$ for all $n \in \mathbb{N}$.
We want to inductively define a sequence $(I_k)_{k\in\mathbb{N}}$ of intervals, $I_k = [a_k, b_k]$ for $k \in \mathbb{N}$, such that

(1) $I_{k+1} \subset I_k$ for each $k \in \mathbb{N}$,

(2) $b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k)$ for each $k \in \mathbb{N}$,

(3) each I_k contains infinitely many elements of $(x_n)_{n\in\mathbb{N}}$.

We put $I_1 := [a_1, b_1]$. Assume that for some $k \in \mathbb{N}$ there are intervals $I_1, \ldots, I_k = [a_k, b_k]$ satisfying these conditions. Let $M := \frac{a_k+b_k}{2}$, i.e. M is the middle point of the interval I_k, and define $I_{k+1} = [a_{k+1}, b_{k+1}]$ by

$$I_{k+1} := \begin{cases} [a_k, M] & \text{if } [a_k, M] \text{ contains infinitely many elements of } (x_n)_{n\in\mathbb{N}}, \\ [M, b_k] & \text{else.} \end{cases}$$

Since I_k contains infinitely many elements of $(x_n)_{n\in\mathbb{N}}$ by assumption and since $[a_k, b_n] = [a_k, M] \cup [M, b_k]$, the interval I_{k+1} is well-defined, satisfies $I_{k+1} \subset I_k$ and contains infinitely many elements of $(x_n)_{n\in\mathbb{N}}$. Moreover, $b_{k+1} - a_{k+1} \in \{M - a_k, b_k - M\}$, but by construction $M - a_k = b_k - M = \frac{1}{2}(b_k - a_k)$. Hence, $b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k)$ as well.
Thus, we obtain a sequence of intervals satisfying the above properties (1), (2) and (3). By induction, one derives from (2) that

$$b_k - a_k = \left(\frac{1}{2}\right)^{k-1}(b_1 - a_1) \qquad \forall k \in \mathbb{N},$$

so it follows that

$$\lim_{k\to\infty}(b_k - a_k) = \left(\lim_{n\to\infty}\left(\frac{1}{2}\right)^{k-1}\right)(b_1 - a_1) = 0$$

as seen in Theorem 2.6. Hence, $(I_k)_{k\in\mathbb{N}}$ satisfies the conditions of the nested interval principle (Theorem 2.47), which implies that there exists $c \in \mathbb{R}$ with $\bigcap_{k\in\mathbb{N}} I_k = \{c\}$.
We want to construct a subsequence of $(x_n)_{n\in\mathbb{N}}$ which converges against c. Put $n_1 := 1$. Assume that for some $k \in \mathbb{N}$ we have found $n_1, \ldots, n_k \in \mathbb{N}$ with

$$n_1 < n_2 < \cdots < n_k,$$

such that $x_{n_j} \in I_j$for each $j \in \{1, 2, \dots, k\}$. Since I_{k+1} contains infinitely many elements of $(x_n)_{n\in\mathbb{N}}$, there exists $n_{k+1} \in \mathbb{N}$ with $n_{k+1} > n_k$, such that $x_{n_{k+1}} \in I_{k+1}$. We choose and fix such an n_{k+1}.
A subsequence $(x_{n_k})_{k\in\mathbb{N}}$ defined in this way satisfies

$$a_k \leq x_{n_k} \leq b_k \qquad \forall k \in \mathbb{N}.$$

Since, as in the proof of the nested interval principle, $\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = c$, it follows from the squeeze theorem (Theorem 2.28) that $(x_{n_k})_{k\in\mathbb{N}}$ is convergent with $\lim_{k\to\infty} x_{n_k} = c$. Hence, we have found a convergent subsequence of $(x_n)_{n\in\mathbb{N}}$. □

The corresponding property of complex sequences can be derived directly from the previous theorem.

Corollary 2.49 (Bolzano-Weierstraß in $\mathbb{C}$)**.** *Every bounded complex sequence has a convergent subsequence.*

Proof. Let $(z_n)_{n\in\mathbb{N}}$ be a bounded complex sequence, let $a_n = \mathrm{Re}(z_n)$ and $b_n = \mathrm{Im}(z_n)$ for all $n \in \mathbb{N}$. Then $(a_n)_{n\in\mathbb{N}}$ is a bounded real sequence, thus, by Theorem 2.48, has a convergent subsequence $(a_{n_k})_{k\in\mathbb{N}}$. Consider the sequence $(b_{n_k})_{k\in\mathbb{N}}$. It is a bounded real sequence and has a convergent subsequence $(b_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ by Theorem 2.48. The sequence $(a_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ now converges by Theorem 2.19. Thus,

$$z_{n_{k_\ell}} = a_{n_{k_\ell}} + i b_{n_{k_\ell}}, \quad \ell \in \mathbb{N},$$

is a subsequence of $(z_n)_{n\in\mathbb{N}}$ which converges by Theorem 2.20. □

Before proceeding with real sequences, we want to fill a gap in Section 1.3. In Theorem 1.48 we claimed the existence of n-th roots of non-negative numbers, but we did not prove it. We want to provide the proof at this point as a consequence of the nested interval principle.

Proof of Theorem 1.48. (This proof is not discussed in the lecture videos.) Throughout the proof, we assume that $a \geq 1$. (The case of $a < 1$ can then be derived from the other case applied to $\frac{1}{a}$.)
We want to inductively define a sequence of closed bounded intervals $(I_k)_{k\in\mathbb{N}}$, $I_k = [a_k, b_k]$ for every $k \in \mathbb{N}$, with the following properties for each $k \in \mathbb{N}$:

(1) $I_{k+1} \subset I_k$,

(2) $a_k^n \leq a \leq b_k^n$,

(3) $b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k)$.

Define $I_1 := [1, a]$. By assumption, $1^n = 1 \leq a \leq a^n$.
Assume that for some $k \in \mathbb{N}$ we have closed bounded intervals $I_1, \dots, I_k$ fulfilling conditions (1), (2) and (3). Let $M := \frac{a_k+b_k}{2}$ be the middle point of I_k and define $I_{k+1} = [a_{k+1}, b_{k+1}]$ by

$$I_{k+1} := \begin{cases} [a_k, M] & \text{if } M^n \geq a, \\ [M, b_k] & \text{if } M^n < a. \end{cases}$$

By construction, I_{k+1} is well-defined and satisfies $I_{k+1} \subset I_k$, $a_{k+1}^n \leq a \leq b_{k+1}^n$ and $b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k)$. As in the proof of the Bolzano-Weierstraß theorem, one uses the nested interval theorem to show that there exists a unique $c \in \mathbb{R}$ with $\bigcap_{k\in\mathbb{N}} I_k = \{c\}$. Since $I_1 \subset [0, +\infty)$, it further holds that $c \geq 0$. We want to show that this c satisfies $c^n = a$.

For each $k \in \mathbb{N}$ we define $I_k^n := [a_k^n, b_k^n]$. Since $(I_k)_{k\in\mathbb{N}}$ satisfies the above conditions, one derives from the above that $(I_k^n)_{k\in\mathbb{N}}$ satisfies $I_{k+1}^n \subset I_k^n$ and that $\lim_{k\to\infty}(b_k^n - a_k^n) = 0$, hence by the nested interval theorem, there exists $c' \geq 0$ with $\bigcap_{k\in\mathbb{N}} I_k^n = \{c'\}$. By condition (2) from above, $a \in I_k^n$ for each $k \in \mathbb{N}$. Moreover, since $c \in I_k$ for each $k \in \mathbb{N}$, it holds that $c^n \in I_k^n$ for each $k \in \mathbb{N}$, so $a, c^n \in \bigcap_{k\in\mathbb{N}} I_k^n$, which shows that $c' = a = c^n$, so c is indeed the unique non-negative solution of $x^n = a$. □

One might encounter sequences whose elements seem to be getting "closer and closer" if n becomes bigger and bigger, while an explicit limit is hard to write down. This is formalized by the notion of Cauchy sequences.

Definition 2.50. Let $K = \mathbb{R}$ or $K = \mathbb{C}$. A sequence $(a_n)_{n\in\mathbb{N}}$ in K is called a *Cauchy sequence* if

$$\forall \varepsilon > 0\ \exists n_0 \in \mathbb{N} :\ |a_n - a_m| < \varepsilon \quad \forall m, n \geq n_0.$$

However, it turns out that the notion of a Cauchy sequence simply recovers the notion of a convergent sequences in $\mathbb{R}$ or $\mathbb{C}$.

Theorem 2.51. *A real sequence is convergent if and only if it is a Cauchy sequence.*

Proof. "⇒": Let $(a_n)_{n\in\mathbb{N}}$ be a convergent real sequence with $a := \lim_{n\to\infty} a_n$. Let $\varepsilon > 0$. Then there exists $n_0 \in \mathbb{N}$ such that $|a_n - a| < \frac{\varepsilon}{2}$ for all $n \geq n_0$, which implies

$$|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a_m - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \forall m, n \geq n_0.$$

Hence, $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence.
"⇐": Let $(a_n)_{n\in\mathbb{N}}$ be a Cauchy sequence. In particular, there exists $n_0 \in \mathbb{N}$ such that

$$|a_n - a_{n_0}| < 1 \qquad \forall n \geq n_0.$$

With the triangle inequality we obtain

$$|a_n| \leq |a_n - a_{n_0}| + |a_{n_0}| \leq 1 + |a_{n_0}| \qquad \forall n \geq n_0,$$

so it follows with $C := \max\{|a_1|, \ldots, |a_{n_0-1}|, 1 + |a_{n_0}|\}$ that $|a_n| \leq C$ for all $n \in \mathbb{N}$. Hence, $(a_n)_{n\in\mathbb{N}}$ is bounded. By the Bolzano-Weierstraß theorem, $(a_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(a_{n_k})_{k\in\mathbb{N}}$. We denote its limit by a.
Let $\varepsilon > 0$. Then there exists $k_0 \in \mathbb{N}$ such that

$$|a_{n_k} - a| < \frac{\varepsilon}{2} \qquad \forall k \geq k_0.$$

Since $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, there further exists $k_1 \in \mathbb{N}$ such that

$$|a_n - a_m| < \frac{\varepsilon}{2} \qquad \forall n, m \geq n_{k_1}.$$

Putting $k_2 := \max\{k_0, k_1\}$, we obtain with the triangle inequality that

$$|a_n - a| = |a_n - a_{n_{k_2}} + a_{n_{k_2}} - a| \leq |a_n - a_{n_{k_2}}| + |a_{n_{k_2}} - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \forall n \geq n_{k_2}.$$

Since ε was arbitrary, this shows that $(a_n)_{n\in\mathbb{N}}$ converges against a. □

Corollary 2.52. *A complex sequence is convergent if and only if it is a Cauchy sequence.*

Proof. (This proof is not discussed in the lecture videos.)
"⇒": If $(z_n)_{n\in\mathbb{N}}$ converges in $\mathbb{C}$, then Theorems 2.20 and 2.51 yield that $(\mathrm{Re}(z_n))_{n\in\mathbb{N}}$ and $(\mathrm{Im}(z_n))_{n\in\mathbb{N}}$ are convergent. Together with the fact that $|z_n - z_m| \leq |\mathrm{Re}(z_n) - \mathrm{Re}(z_m)| + |\mathrm{Im}(z_n) - \mathrm{Im}(z_m)|$, this implies that $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence.
"⇐": If $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}$, it is easy to see that both $(\mathrm{Re}(z_n))_{n\in\mathbb{N}}$ and $(\mathrm{Im}(z_n))_{n\in\mathbb{N}}$ are Cauchy sequences in $\mathbb{R}$. The claim then follows from Theorems 2.51 and 2.20. □

Remark 2.53. (Just for those who are interested, you may skip this remark without disadvantages.) The convergence of Cauchy sequences in $\mathbb{R}$ is intimately related to the completeness of $\mathbb{R}$. In fact, one might reformulate the definition of a complete ordered by axiomatically demanding the convergence of Cauchy sequences and the unboundedness of $\mathbb{N}$.

Definition 2.54. Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K. $a \in K$ is called an *accumulation point of* $(a_n)_{n\in\mathbb{N}}$ if for each $\varepsilon > 0$ the set $\{n \in \mathbb{N} \mid |x_n - a| < \varepsilon\}$ is infinite.

Theorem 2.55. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $a \in K$. The following statements are equivalent:*

(i) a is an accumulation point of $(a_n)_{n\in\mathbb{N}}$.

(ii) $\forall \varepsilon > 0 \ \forall N \in \mathbb{N} \ \exists n \in \mathbb{N}$ with $n \geq N$: $\quad |a_n - a| < \varepsilon$.

(iii) There exists a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ of $(a_n)_{n\in\mathbb{N}}$ with $\lim_{k\to\infty} a_{n_k} = a$.

Remark 2.56. A general remark on the strategy of proof for Theorem 2.55:
To show the equivalents of more than two statements it is not necessary to individually prove that any two of the statements are equivalent. If the statements are given by $A_1, \ldots, A_n$, it suffices to prove the implications

$$A_1 \Rightarrow A_2, \quad A_2 \Rightarrow A_3, \quad \ldots, \quad A_{n-1} \Rightarrow A_n, \quad A_n \Rightarrow A_1, \tag{2.4}$$

or, alternatively, the implications

$$A_n \Rightarrow A_{n-1}, \quad A_{n-1} \Rightarrow A_{n-2}, \quad \ldots, \quad A_2 \Rightarrow A_1, \quad A_n \Rightarrow A_1.$$

Then all other equivalences follow: if the equivalences in (2.4) are shown and if $i, j \in \{1, 2, \ldots n\}$ with $i < j$, then the chain of implications $A_i \Rightarrow A_{i+1} \Rightarrow \cdots \Rightarrow A_j$ shows that $A_i \Rightarrow A_j$, while the chain of implications $A_j \Rightarrow A_{j+1} \Rightarrow \cdots \Rightarrow A_n \Rightarrow A_1 \Rightarrow A_2 \Rightarrow \cdots \Rightarrow A_i$ shows that $A_j \Rightarrow A_i$.

Proof of Theorem 2.55. (i) ⇒ (ii): If a is an accumulation point of $(a_n)_{n\in\mathbb{N}}$, then for each $\varepsilon > 0$ the set $\{n \in \mathbb{N} \mid |a_n - a| < \varepsilon\}$ is infinite. Consequently, the set $\{n \in \mathbb{N} \mid n \geq N, \ |a_n - a| < \varepsilon\}$ is non-empty for each $N \in \mathbb{N}$, which shows the claim.
(ii) ⇒ (iii): We inductively construct a subsequence as follows:
Pick $n_1 \in \mathbb{N}$, such that $|a_{n_1} - a| < 1$, which exists by assumption. Given $n_1 < n_2 < \cdots < n_k$, let $n_{k+1} \in \mathbb{N}$ be given such that

$$n_{k+1} > n_k, \qquad |a_{n_{k+1}} - a| < \frac{1}{k+1}.$$

Such an n_{k+1} exists by assumption (ii) with $\varepsilon = \frac{1}{k+1}$ and $N = n_k + 1$. The subsequence $(a_{n_k})_{k\in\mathbb{N}}$ then satisfies

$$0 \leq |a_{n_k} - a| \leq \frac{1}{k} \quad \forall k \in \mathbb{N},$$

so by the squeeze theorem, more precisely by Corollary 2.29, it follows that $\lim_{k\to\infty} |a_{n_k} - a| = 0$, which implies $\lim_{k\to\infty} a_{n_k} = a$.
(iii) $\Rightarrow$ (i): Let $(a_{n_k})_{k\in\mathbb{N}}$ be a subsequence with $\lim_{k\to\infty} a_{n_k} = a$ and let $\varepsilon > 0$. Then there exists $k_0 \in \mathbb{N}$ with $|a_{n_k} - a| < \varepsilon$ for all $k \geq k_0$. Since $\{n_k \mid k \geq k_0\} \subset \{n \in \mathbb{N} \mid |a_n - a| < \varepsilon\}$, it follows that the latter set is infinite. Since $\varepsilon > 0$ was arbitrary, a is an accumulation point of $(a_n)_{n\in\mathbb{N}}$. □

Remark 2.57. Combining the Bolzano-Weierstraß theorem with Theorem 2.55 shows that every bounded real or complex sequence has an accumulation point.

Example 2.58. (1) Let $a_n = (-1)^n + \frac{1}{n}$, $n \in \mathbb{N}$. Then $(a_n)_{n\in\mathbb{N}}$ has the accumulation points -1 and 1, since

$$\lim_{k\to\infty} a_{2k} = \lim_{k\to\infty} (1 + \frac{1}{2k}) = 1,$$
$$\lim_{k\to\infty} a_{2k+1} = \lim_{k\to\infty} (-1 + \frac{1}{2k+1}) = -1.$$

Moreover, one shows that for any convergent subsequence of $(a_n)_{n\in\mathbb{N}}$ either almost every element of the subsequence lies in $(a_{2k})_{k\in\mathbb{N}}$ or almost every element lies in $(a_{2k+1})_{k\in\mathbb{N}}$, so there can be no more accumulation points.

(2) $a_n = n$, $n \in \mathbb{N}$, has *no* accumulation point since every subsequence of $(a_n)_{n\in\mathbb{N}}$ is unbounded, hence divergent by Theorem 2.10.

(3) Let

$$a_n = \begin{cases} n & \text{if } n \text{ is even,} \\ \frac{1}{n} & \text{if } n \text{ is odd.} \end{cases}$$

Then $(a_n)_{n\in\mathbb{N}}$ is not bounded from above, hence divergent, but it has the accumulation point 0, since

$$\lim_{k\to\infty} a_{2k+1} = \lim_{k\to\infty} \frac{1}{2k+1} = 0.$$

Definition 2.59. Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence.

a) If $(a_n)_{n\in\mathbb{N}}$ is bounded from above, then we define the *limit superior of* $(a_n)_{n\in\mathbb{N}}$ by

$$\limsup_{n\to\infty} a_n := \sup\{a \in \mathbb{R} \mid a \text{ is an accumulation point of } (a_n)_{n\in\mathbb{N}}\}.$$

If $(a_n)_{n\in\mathbb{N}}$ is not bounded from above, then we put $\limsup_{n\to\infty} a_n := +\infty$. We also write: $\overline{\lim}_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$.

b) If $(a_n)_{n\in\mathbb{N}}$ is bounded from below, then we define the *limit inferior of* $(a_n)_{n\in\mathbb{N}}$ by

$$\liminf_{n\to\infty} a_n := \inf\{a \in \mathbb{R} \mid a \text{ is an accumulation point of } (a_n)_{n\in\mathbb{N}}\}.$$

If $(a_n)_{n\in\mathbb{N}}$ is not bounded from below, then we put $\liminf_{n\to\infty} a_n := -\infty$. We also write: $\underline{\lim}_{n\to\infty} a_n = \liminf_{n\to\infty} a_n$.

Theorem 2.60. *Let $(a_n)_{n\in\infty}$ be a bounded real sequence. Then $\overline{\lim}_{n\to\infty} a_n$ and $\underline{\lim}_{n\to\infty} a_n$ are themselves accumulation points of $(a_n)_{n\in\mathbb{N}}$.*

Proof. We will only show that $\overline{\lim}_{n\to\infty} a_n$ is an accumulation point of $(a_n)_{n\in\mathbb{N}}$, since the statement for $\underline{\lim}_{n\to\infty} a_n$ is proven analogously.
Put $S := \overline{\lim}_{n\in\mathbb{N}} a_n$ and let $\varepsilon > 0$. By definition of S as a supremum, it follows that there exists an accumulation point a of $(a_n)_{n\in\mathbb{N}}$ with

$$S - \frac{\varepsilon}{2} < a \leq S.$$

Since a is an accumulation point, the set $M := \{n \in \mathbb{N} \mid |a_n - a| < \frac{\varepsilon}{2}\}$ is infinite. Since

$$|a_n - S| \leq |a_n - a| + |a - S| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \forall n \in M,$$

it follows that $M \subset \{n \in \mathbb{N} \mid |a_n - S| < \varepsilon\}$, so the latter set is infinite as well. Since $\varepsilon > 0$ was arbitrary, S is an accumulation point of $(a_n)_{n\in\mathbb{N}}$. □

The following theorem shows that for n becoming bigger, every bounded sequence "scatters" close to the interval $[\underline{\lim}_{n\to\infty} a_n, \overline{\lim}_{n\to\infty} a_n]$.

Theorem 2.61. *Let $(a_n)_{n\in\mathbb{N}}$ be a bounded real sequence. Then*

$$\forall \varepsilon > 0\ \exists n_0 \in \mathbb{N}, \text{ such that } a_n \in \Big(\underline{\lim}_{n\to\infty} a_n - \varepsilon, \overline{\lim}_{n\to\infty} a_n + \varepsilon \Big) \quad \forall n \geq n_0.$$

Proof. Let $\varepsilon > 0$ and consider the sets

$$M_1 := \{n \in \mathbb{N} \mid a_n \geq \overline{\lim}_{n\to\infty} a_n + \varepsilon\}, \qquad M_2 := \{n \in \mathbb{N} \mid a_n \leq \underline{\lim}_{n\to\infty} a_n - \varepsilon\}.$$

Assume that M_1 is infinite. Then there is a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ with $n_k \in M_1$ for each $k \in \mathbb{N}$. Since $(a_{n_k})_{k\in\mathbb{N}}$ is bounded, it has a convergent subsequence $(a_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ by the Bolzano-Weierstraß Theorem. Since $(a_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ is also a convergent subsequence of $(a_n)_{n\in\mathbb{N}}$, it follows from Theorem 2.55 that $a := \lim_{\ell\to\infty} a_{n_{k_\ell}}$ is an accumulation point of $(a_n)_{n\in\mathbb{N}}$, hence $a \leq \overline{\lim} a_n$, But this is a contradiction since $a \geq \overline{\lim} a_n + \varepsilon$ by Corollary 2.27.
Hence, M_1 is finite. Analogougsly, one shows that M_2 is finite. Hence, $n_0 := \max(M_1 \cup M_2) \in \mathbb{N}$ is well-defined and it holds that

$$a_n \in \Big(\underline{\lim}_{n\to\infty} a_n - \varepsilon, \overline{\lim}_{n\to\infty} a_n + \varepsilon \Big) \qquad \forall n \geq n_0.$$

□

Corollary 2.62. *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence and $a \in \mathbb{R}$. Then $(a_n)_{n\in\mathbb{N}}$ converges to a if and only if $(a_n)_{n\in\mathbb{N}}$ is bounded and $\overline{\lim}_{n\to\infty} a_n = \underline{\lim}_{n\to\infty} a_n = a$.*

Proof. "⇐": This follows immediately from Theorem 2.61.
"⇒": By Theorem 2.10, $(a_n)_{n\in\mathbb{N}}$ is bounded. By Theorem 2.55, a is an accumulation point of $(a_n)_{n\in\mathbb{N}}$. By Theorem 2.19 and Theorem 2.55 it is the only accumulation point of $(a_n)_{n\in\mathbb{N}}$. Hence, $\overline{\lim}_{n\to\infty} a_n = \underline{\lim}_{n\to\infty} a_n = a$. □

2.4 Series and convergence tests

Imagine that a liquid is running into a certain container, but pressure is dropping so that every minute less and less liquid is running into the container. Let a_n be the amound of liquid running into the container in minute n. Then the total amount of liquid in the container after n minutes is given by

$$a_1 + a_2 + a_3 + \cdots + a_n.$$

Will the container flow over at some point? Will the stream of liquid become so thin that the container is big enough for all of the liquid pouring in? To answer these questions, we need to study the above sum for n becoming bigger and bigger. Intuitively, we have to study an "infinite sum".

The notion of a *series* formalizes the notion of an "infinite sum" by adding up elements of sequences of real or complex numbers. For example, one might ask what happens if we add more and more terms in the following sums:

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \cdots = ?$$
$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots = ?$$

Do these sums become infinitely big? Will they stay below a certain upper bound? We can imagine these infinite sums as limits of sequences. We start with a certain number (in both examples with 1) and start adding elements by a fixed scheme, in a fixed order. The n-th element of such a sequence would be the sum of the first n terms. While it turns out that the limit of such series are in general hard to compute, it turns out that there are various methods to make statements about the convergence of series as we shall see in the course of this section.

Definition 2.63. Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K. The sequence $(s_n)_{n\in\mathbb{N}}$, where

$$s_n = \sum_{k=1}^{n} a_k \qquad \forall n \in \mathbb{N},$$

is called the *series of* $(a_n)_{n\in\mathbb{N}}$ and is denoted by $\sum_{n=1}^{\infty} a_n$. The term s_n is called the *n-th partial sum of the series*. We say that $\sum_{n=1}^{\infty} a_n$ is convergent (divergent) if $(s_n)_{n\in\mathbb{N}}$ is convergent (divergent).

If the series $\sum_{n=1}^{\infty} a_n$ converges, then we put

$$\sum_{n=1}^{\infty} a_n := \lim_{n\to\infty} s_n.$$

If $K = \mathbb{R}$ we call $\sum_{n=1}^{\infty} a_n$ a *real series*, if $K = \mathbb{C}$ we call it a *complex series*.

For technical reasons one is often interested in sums which do not necessarily start at $n = 1$, but at some number $n = n_0 \in \mathbb{Z}$. To include these cases in our studies of convergence, we introduce some additional formalism.

Definition 2.64. Let $n_0 \in \mathbb{Z}$ be a given number and let $(a_n)_{n \geq n_0}$ be a family of complex numbers. We define $\sum_{n=n_0}^{\infty} a_n$ as the series $\sum_{n=1}^{\infty} a_{n+n_0-1}$, i.e. as the series whose n-th partial sum is given by

$$s_n = \sum_{k=1}^{n} a_{k+n_0-1} = \sum_{k=n_0}^{n_0+n-1} a_k.$$

We say that $\sum_{n=n_0}^{\infty} a_n$ is *convergent* (or *divergent*) if the series $\sum_{n=1}^{\infty} a_{n+n_0-1}$ is convergent (or divergent).

Example 2.65. (1) The series

$$\sum_{n=0}^{\infty} q^n, \qquad q \in \mathbb{C}.$$

is called *the geometric series in q*. Let $s_n = \sum_{k=1}^{n} q^k$ be its n-th partial sum. In Exercise 2b) of Sheet 2 it was shown that

$$s_n = \frac{1-q^{n+1}}{1-q} \qquad \forall n \in \mathbb{N}.$$

for each $q \in \mathbb{C}$ with $q \neq 1$. Thus, we derive from Theorem 2.6 and the computational rules for limits that

$$\sum_{n=0}^{\infty} q^n = \lim_{n\to\infty} \frac{1-q^{n+1}}{1-q} = \frac{1-\lim_{n\to\infty} q^{n+1}}{1-q} = \frac{1}{1-q}.$$

For example, for $q = \frac{1}{2}$ and $q = -\frac{1}{2}$ this implies that

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots = \sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{1-\frac{1}{2}} = 2,$$
$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \ldots = \sum_{n=1}^{\infty} \Big(-\frac{1}{2}\Big)^n = \frac{1}{1+\frac{1}{2}} = \frac{2}{3}.$$

(2) We consider the series $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$. We compute for each $n \in \mathbb{N}$ that

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

Hence, with s_n denoting the n-th partial sum, we obtain

$$s_n = \sum_{k=1}^{n} \Big(\frac{1}{k} - \frac{1}{k+1}\Big) = \Big(1 - \frac{1}{2}\Big) + \Big(\frac{1}{2} - \frac{1}{3}\Big) + \cdots + \Big(\frac{1}{n} - \frac{1}{n+1}\Big) = 1 - \frac{1}{n+1}.$$

Thus,

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{n\to\infty} \Big(1 - \frac{1}{n+1}\Big) = 1.$$

Theorem 2.66. *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with $a_n \geq 0$ for all $n \in \mathbb{N}$. Then $\sum_{n=1}^{\infty} a_n$ is convergent if and only if it is bounded from above. Otherwise, it diverges to $+\infty$.*

Proof. Any convergent sequence is bounded, so we only need to study the converse statement. This follows from Theorem 2.35 since $a_n \geq 0$ for all n implies that the sequence of partial sums is monotonically increasing. □

Example 2.67. The *harmonic series* is the series

$$\sum_{n=1}^{\infty} \frac{1}{n}.$$

Let s_n denote its n-th partial sum. We want to show that $(s_n)_{n\in\mathbb{N}}$ has an unbounded subsequence, which implies by Theorem 2.66 that the series diverges to $+\infty$.
Given $n \in \mathbb{N}$ we consider:

$$\begin{aligned} s_{2^k} &= \sum_{n=1}^{2^k} \frac{1}{n} = 1 + \frac{1}{2} + \sum_{i=1}^{k-1} \Big(\sum_{n=2^i+1}^{2^{i+1}} \frac{1}{n} \Big) \\ &= 1 + \frac{1}{2} + \Big(\frac{1}{3} + \frac{1}{4}\Big) + \Big(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\Big) + \cdots + \Big(\frac{1}{2^{k-1}+1} + \cdots + \frac{1}{2^k}\Big). \end{aligned}$$

For each $i \in \{1, \ldots, k-1\}$ we compute that

$$\sum_{n=2^i+1}^{2^{i+1}} \frac{1}{n} \geq \sum_{n=2^i+1}^{2^{i+1}} \frac{1}{2^{i+1}} = 2^i \cdot \frac{1}{2^{i+1}} = \frac{1}{2}.$$

Hence,

$$s_{2^k} \geq 1 + \frac{1}{2} + \sum_{i=1}^{k-1} \frac{1}{2} = 1 + \frac{k}{2}.$$

Thus, by Theorem 2.42, $(s_{2^k})_{k\in\mathbb{N}}$ diverges to $+\infty$ and is unbounded. It follows from Theorem 2.66 that the harmonic series diverges to $+\infty$.

In the rest of this section, we want to investigate the convergence of series in greater detail. We first establish a *necessary* condition for a series to converge.

Theorem 2.68. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K. If the series $\sum_{n=1}^{\infty} a_n$ converges, then $\lim_{n\to\infty} a_n = 0$.*

Proof. Let s_n be the n-th partial sum of $\sum_{n=1}^{\infty} a_n$. Since $(s_{n+1})_{n\in\mathbb{N}}$ is a subsequence of $(s_n)_{n\in\mathbb{N}}$, it converges against the same limit by Theorem 2.19. Hence,

$$0 = \lim_{n\to\infty} s_{n+1} - \lim_{n\to\infty} s_n = \lim_{n\to\infty} (s_{n+1} - s_n) = \lim_{n\to\infty} \Big(\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k \Big) = \lim_{n\to\infty} a_{n+1}.$$

Hence, $\lim_{n\to\infty} a_n = 0$. □

Remark 2.69. The converse statement of Theorem 2.68 does not hold. A counterexample is given by the harmonic series. Here, $\lim_{n\to\infty} \frac{1}{n} = 0$, but the series diverges to $+\infty$.

Theorem 2.70. *Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ be convergent series in K with limits $\sum_{n=1}^{\infty} a_n = a$ and $\sum_{n=1}^{\infty} b_n = b$.*

a) $\sum_{n=1}^{\infty}(a_n + b_n)$ is convergent with $\sum_{n=1}^{\infty}(a_n + b_n) = \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n = a + b$.

b) For each $c \in K$ the series $\sum_{n=1}^{\infty}(ca_n)$ converges with $\sum_{n=1}^{\infty}(ca_n) = c\sum_{n=1}^{\infty} a_n = ca$.

Proof. This is a special case of Theorem 2.7 applied to sequences of partial sums. □

... Series are special cases of sequences and using the ... we want to establish a number of *convergence tests*, i.e. a number of sufficient criteria that imply the convergence of a series.

Definition 2.71. A real series is called *alternating* if it is of the form

$$\sum_{n=1}^{\infty}(-1)^n a_n \qquad \text{or} \qquad \sum_{n=1}^{\infty}(-1)^{n+1} a_n,$$

where $(a_n)_{n\in\mathbb{N}}$ is a real sequence with $a_n \geq 0$ for each $n \in \mathbb{N}$.

Theorem 2.72 (Leibniz's convergence test)**.** *Let $\sum_{n=1}^{\infty}(-1)^{n+1} a_n$ be an alternating series. If $(a_n)_{n\in\mathbb{N}}$ is monotonically decreasing with $\lim_{n\to\infty} a_n = 0$, then $\sum_{n=1}^{\infty}(-1)^n a_n$ is convergent.*

Proof. Let $s_n = \sum_{k=1}^{n}(-1)^{k+1} a_k$ be the n-th partial sum of the series and consider the subsequences $(s_{2k})_{k\in\mathbb{N}}$ and $(s_{2k+1})_{k\in\mathbb{N}}$ of $(s_n)_{n\in\mathbb{N}}$. For each $k \in \mathbb{N}$ we obtain using the monotonicity of $(a_n)_{n\in\mathbb{N}}$ that

$$s_{2k+2} = s_{2k} + \underbrace{(a_{2k+1} - a_{2k+2})}_{\geq 0} \geq s_{2k}$$

and

$$s_{2k+3} = s_{2k+1} + \underbrace{(-a_{2k+2} + a_{2k+3})}_{\leq 0} \leq s_{2k+1}.$$

Hence, $(s_{2k})_{k\in\mathbb{N}}$ is monotonically increasing while (s_{2k+1}) is monotonically decreasing. Moreover,

$$s_{2k+1} - s_{2k} = a_{2k+1} \geq 0 \quad \forall k \in \mathbb{N}.$$

Together with the monotonicities this shows that $s_{2k} \leq s_{2k+1} \leq s_3$ and $s_{2k+1} \geq s_{2k} \geq s_2$, such that $(s_{2k})_{k\in\mathbb{N}}$ is bounded from above while $(s_{2k+1})_{k\in\mathbb{N}}$ is bounded from below. Thus, both subsequence converge by Theorem 2.35. We further derive that

$$\lim_{k\to\infty} s_{2k+1} - \lim_{k\to\infty} s_{2k} = \lim_{k\to\infty}(s_{2k+1} - s_{2k}) = \lim_{k\to\infty} a_{2k+1} = 0$$

by assumption. Hence, both subsequences have the same limit L. Let $\varepsilon > 0$. Then there exist $k_1, k_2 \in \mathbb{N}$ such that

$$|s_{2k} - L| < \varepsilon \quad \forall k \geq k_1, \qquad |s_{2k+1} - L| < \varepsilon \quad \forall k \geq k_2.$$

Putting $n_0 := \max\{2k_1, 2k_2 + 1\}$, we obtain that

$$|s_n - L| < \varepsilon \quad \forall n \geq n_0.$$

This shows that $\lim_{n\to\infty} s_n = L$, hence $\sum_{n=1}^{\infty}(-1)^{n+1}a_n$ converges. □

Example 2.73. By Leibniz's convergence test, the *alternating harmonic series* $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ is convergent. However, the Leibniz test offers no possibility of computing its limit. (Later, using some powerful tools from calculus, it will turn out that $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n}{n} = \log 2$.)

Throughout the following, we consider $K = \mathbb{R}$ or $K = \mathbb{C}$ and let all sequences and series be sequences and series in K.

Theorem 2.74 (Cauchy's convergence test)**.** *The series $\sum_{n=1}^{\infty} a_n$ converges if and only if*

$$\forall \varepsilon > 0\ \exists n_0 \in \mathbb{N}: \quad \left| \sum_{k=m+1}^{n} a_k \right| < \varepsilon \quad \forall n > m \geq n_0.$$

Proof. By Theorem 2.51, the series converges if and only if the sequence $s_n = \sum_{k=1}^{n} a_k$, $n \in \mathbb{N}$, is a Cauchy sequence. Since for all $n > m$ it holds that

$$|s_n - s_m| = \left| \sum_{k=1}^{n} a_k - \sum_{k=1}^{m} a_k \right| = \left| \sum_{k=m+1}^{n} a_k \right|,$$

the claim follows from the definition of a Cauchy sequence. □

The following notion is a stronger version of convergence that may not look particularly interesting at this point, but which turns out to have powerful consequences later on.

Definition 2.75. A series $\displaystyle\sum_{n=1}^{\infty} a_n$ is called *absolutely convergent* if the series $\displaystyle\sum_{n=1}^{\infty} |a_n|$ is convergent.

Let us clarify the relation between convergence and absolute convergence of a series.

Theorem 2.76. *Every absolutely convergent series is convergent.*

Proof. Let $\sum_{n=1}^{\infty} a_n$ be absolutely convergent. Let $\varepsilon > 0$. By Theorem 2.74, there exists $n_0 \in \mathbb{N}$, such that

$$\sum_{k=m+1}^{n} |a_k| = \left| \sum_{k=m+1}^{n} |a_k| \right| < \varepsilon \quad \forall n > m \geq n_0.$$

By the triangle inequality, this implies

$$\left| \sum_{k=m+1}^{n} a_k \right| \leq \sum_{k=m+1}^{n} |a_k| < \varepsilon \quad \forall n > m \geq n_0.$$

Since $\varepsilon > 0$ was arbitrary, Theorem 2.74 implies that $\sum_{n=1}^{\infty} a_n$ is convergent. □

Remark 2.77. The converse statement to Theorem 2.76 is false. A counterexample is given by $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$, as we have seen in Examples 2.67 and 2.73.

Theorem 2.78 (Comparison test). *Let* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *be series with* $|a_n| \leq |b_n|$ *for almost every* $n \in \mathbb{N}$.

a) If $\sum_{n=1}^{\infty} b_n$ *converges absolutely, then* $\sum_{n=1}^{\infty} a_n$ *converges absolutely.*

b) If $\sum_{n=1}^{\infty} a_n$ *diverges, then* $\sum_{n=1}^{\infty} |b_n|$ *diverges.*

Proof. a) Let $n_0 \in \mathbb{N}$ with $|a_n| \leq |b_n|$ for all $n \geq n_0$. It then holds for all $n > n_0$ that

$$\sum_{k=1}^{n} |a_k| = \sum_{k=1}^{n_0} |a_k| + \sum_{k=n_0+1}^{n} |a_k| \leq \sum_{k=1}^{n_0} |a_k| + \sum_{k=n_0+1}^{n} |b_k| \leq \sum_{k=1}^{n_0} |a_k| + \sum_{n=1}^{\infty} |b_n| =: C.$$

Since $\sum_{n=1}^{\infty} b_n$ converges absolutely, the upper bound $C \in \mathbb{R}$ is well-defined. Thus, the sequence of partial sums of $\sum_{n=1}^{\infty} |a_n|$ is bounded from above, thus convergent by Theorem 2.66. This shows the absolute convergence of $\sum_{n=1}^{\infty} a_n$.

b) Since $\sum_{n=1}^{\infty} a_n$ diverges, Theorem 2.76 implies that it does not converge absolutely. Hence, by part a), $\sum_{n=1}^{\infty}$ can not converge absolutely.

□

Example 2.79. We consider the series $\sum_{n=1}^{\infty} \frac{1}{n^k}$, where $k \geq 2$. We compute that

$$\frac{1}{n^k} \leq \frac{1}{n^2} \leq \frac{2}{n(n+1)} \qquad \forall n \in \mathbb{N},$$

where we have used that $n^2 = \frac{2n^2}{2} \geq \frac{n^2+n}{2} = \frac{n(n+1)}{2}$ for each $n \in \mathbb{N}$. By Example 2.65.(2), it holds that $\sum_{n=1}^{\infty} \frac{2}{n(n+1)} = 2\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 2$. Hence, by part a) of the comparison test, $\sum_{n=1}^{\infty} \frac{1}{n^k}$ converges absolutely for all $k \geq 2$.
However, it turns out that the actual limits of these series are very hard to compute. For example, with a lot of additional effort, one can prove that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, which is totally out of reach with our current knowledge.

Theorem 2.80 (Root test). *Let* $\sum_{n=1}^{\infty} a_n$ *be a series and put*

$$r := \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

a) If $r < 1$, *then* $\sum_{n=1}^{\infty} a_n$ *converges absolutely.*

b) If $r > 1$, *then* $\sum_{n=1}^{\infty} a_n$ *is divergent.*

Proof. a) Let $q \in (r,1)$. By Theorem 2.61, there exists $n_0 \in \mathbb{N}$, such that $\sqrt[n]{|a_n|} \leq q$ and thus $|a_n| \leq q^n$ for all $n \geq n_0$. Thus, applying the comparison test with $b_n = q^n$, it follows from Theorem 2.78.a) that $\sum_{n=1}^{\infty} a_n$ converges absolutely.

b) If $\limsup_{n\to\infty} \sqrt[n]{|a_n|} = +\infty$, then $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$ is unbounded, hence $(|a_n|)_{n\in\mathbb{N}}$ is unbounded, such that $(a_n)_{n\in\mathbb{N}}$ is divergent. In particular, Theorem 2.68 then implies that $\sum_{n=1}^{\infty} a_n$ is divergent.

Assume that $r = \limsup_{n\to\infty} \sqrt[n]{|a_n|} \in (1,+\infty)$, such that $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$ is bounded. By Theorem 2.60, r is an accumulation point of $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$. Since $r > 1$, it follows that there is a subsequence $(\sqrt[n_k]{|a_{n_k}|})_{k\in\mathbb{N}}$ with $\sqrt[n_k]{|a_{n_k}|} \geq 1$ and thus $|a_{n_k}| \geq 1$ for all $k \in \mathbb{N}$. This particularly implies that $(a_n)_{n\to\infty}$ does not converge to 0, so by Theorem 2.68 the series $\sum_{n=1}^{\infty} a_n$ diverges.

□

Remark 2.81. If $r = 1$ in the situation of the root test, then it is impossible to make a general convergence statement. For example, for both $\sum_{n=1}^{\infty} \frac{1}{n}$ and $\frac{1}{n^2}$ we obtain $r = 1$ in the root test, but as we have seen the former series diverges, while the latter series converges (absolutely).

Since the sequence $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$ used in the root test might be hard to handle, we study another convergence test which is sometimes easier to compute.

Theorem 2.82 (Ratio test)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence with $a_n \neq 0$ for all $n \in \mathbb{N}$*

a) If

$$r := \limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} < 1,$$

then $\sum_{n=1}^{\infty} a_n$ converges absolutely.

b) If $\frac{|a_{n+1}|}{|a_n|} \geq 1$ for almost every $n \in \mathbb{N}$, then $\sum_{n=1}^{\infty} a_n$ is divergent.

Proof. a) In analogy with the root test, for any $q \in (r,1)$ we obtain an $n_0 \in \mathbb{N}$ with

$$\frac{|a_{n+1}|}{|a_n|} \leq q \quad \forall n \geq n_0 \qquad \Leftrightarrow \qquad |a_{n+1}| \leq q|a_n| \quad \forall n \geq n_0.$$

Inductively, one derives that

$$|a_n| \leq q^{n-n_0}|a_{n_0}| \quad \forall n \geq n_0.$$

Putting $b_n := q^{n-n_0}|a_{n_0}|$ for each $n \in \mathbb{N}$, we obtain that $|a_n| \leq |b_n|$ for almost every $n \in \mathbb{N}$ and that $\sum_{n=1}^{\infty} b_n = q^{-n_0}|a_{n_0}| \sum_{n=1}^{\infty} q^n$ converges by Example 2.65.(1). Hence, $\sum_{n=1}^{\infty} a_n$ converges absolutely by the comparison test.

b) If $|a_{n+1}| \geq |a_n|$ for almost every $n \in \mathbb{N}$, then there exists $n_0 \in \mathbb{N}$ with $|a_n| \geq |a_{n_0}|$ for all $n \geq n_0$. In particular, $(a_n)_{n\in\mathbb{N}}$ does not converge to 0 since $|a_{n_0}| \neq 0$, hence $\sum_{n=1}^{\infty} a_n$ diverges by Theorem 2.68.

□

Remark 2.83. (1) The same examples as in Remark 2.81 show that in the case $r = 1$ in Theorem 2.82, one can not make a general statement about convergence. Hence, same as the root test, the ratio test is a sufficient, but not a necessary condition for convergence.

(2) In the situation of the ratio test, $\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} > 1$ does *not* imply divergence of the series. A counterexample is given by $\sum_{n=1}^{\infty} a_n$, where

$$a_n = \begin{cases} (\frac{1}{2})^n & \text{if } n \text{ is even,} \\ (\frac{3}{4})^n & \text{if } n \text{ is odd.} \end{cases}$$

Convergence of the sequence follows by comparison with $\sum_{n=1}^{\infty}(\frac{3}{4})^n$. But $\left(\frac{|a_{2n+1}|}{|a_{2n}|}\right)_{n\in\mathbb{N}}$ is unbounded, hence $\limsup_{n\to\infty} \frac{|a_{2n+1}|}{|a_{2n}|} = +\infty$.

(3) Every sequence obeying the convergence condition of the ratio test obeys the convergence condition of the root test as well, since one can indeed show that

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} \leq \limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}.$$

Hence, the root test is stronger than the ratio test. (On the other hand, the ratio test is often easier to check than the root test, so both tests have their purposes.)

To conclude this section, we want to discuss an important property of products of series.

Theorem 2.84. *Let $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ be absolutely convergent series. Then the series*

$$\sum_{n=0}^{\infty} c_n, \qquad \text{where} \quad c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad \forall n \in \mathbb{N}_0,$$

is absolutely convergent with

$$\sum_{n=0}^{\infty} c_n = \Big(\sum_{n=0}^{\infty} a_n\Big) \cdot \Big(\sum_{n=0}^{\infty} b_n\Big).$$

Proof. (This proof is not discussed in the lecture videos.)

Let $s_n = \sum_{k=0}^{n} a_k$, $t_n = \sum_{k=0}^{n} b_k$ and $u_n = \sum_{k=0}^{n} c_k$ be the respective partial sum for each $n \in \mathbb{N}$. We first want to show that $(u_n)_{n\in\mathbb{N}}$ converges with

$$\lim_{n\to\infty} s_n \cdot \lim_{n\to\infty} t_n \overset{!}{=} \lim_{n\to\infty} u_n.$$

For this purpose, we first compute for arbitrary $n \in \mathbb{N}$:

$$u_n = \sum_{k=0}^{n} c_k = \sum_{k=0}^{n}\sum_{\ell=0}^{k} a_\ell b_{k-\ell} = \sum_{\ell=0}^{n}\sum_{k=\ell}^{n} a_\ell b_{k-\ell} = \sum_{k=0}^{n}\sum_{\ell=k}^{n} a_k b_{\ell-k}.$$

where we simply switched the names of the indices in the final step. We then perform an index shift to ℓ and obtain

$$u_n = \sum_{k=0}^{n}\sum_{\ell=0}^{n-k} a_k b_\ell.$$

Using this equation, we derive for each $n \in \mathbb{N}$ that

$$\begin{aligned}|s_n \cdot t_n - u_n| &= \left|\sum_{k=0}^{n} a_k \cdot \sum_{\ell=0}^{n} b_\ell - \sum_{k=0}^{n}\sum_{\ell=0}^{n-k} a_k b_\ell\right| \\ &= \left|\sum_{k=0}^{n}\sum_{\ell=0}^{n} a_k b_\ell - \sum_{k=0}^{n}\sum_{\ell=0}^{n-k} a_k b_\ell\right| \\ &= \left|\sum_{k=0}^{n}\sum_{\ell=n-k+1}^{n} a_k b_\ell\right| \leq \sum_{k=0}^{n}\sum_{\ell=n-k+1}^{n} |a_k| \cdot |b_\ell|,\end{aligned}$$

where we used the triangle inequality in the final step. Let $m, n \in \mathbb{N}$ with $n \geq 2m$. Then we compute by first decomposing the sum in the previous estimate and then adding additional summands that

$$\begin{aligned}|s_n \cdot t_n - u_n| &\leq \sum_{k=0}^{m}\sum_{\ell=n-k-1}^{n} |a_k||b_\ell| + \sum_{k=m+1}^{n}\sum_{\ell=n-k-1}^{n} |a_k||b_\ell| \\ &\leq \sum_{k=0}^{m}\sum_{\ell=m+1}^{n} |a_k||b_\ell| + \sum_{k=m+1}^{n}\sum_{\ell=0}^{n} |a_k||b_\ell| \\ &= \sum_{k=0}^{m}\sum_{\ell=0}^{n} |a_k||b_\ell| - \sum_{k=0}^{m}\sum_{\ell=0}^{m} |a_k||b_\ell| + \sum_{k=m+1}^{n}\sum_{\ell=0}^{n} |a_k||b_\ell| \\ &= \sum_{k=0}^{n}\sum_{\ell=0}^{n} |a_k||b_\ell| - \sum_{k=0}^{m}\sum_{\ell=0}^{m} |a_k||b_\ell| \\ &= \bar{s}_n \bar{t}_n - \bar{s}_m \bar{t}_m = |\bar{s}_n \bar{t}_n - \bar{s}_m \bar{t}_m|.\end{aligned}$$

where $\bar{s}_n := \sum_{k=0}^{n} |a_k|$ and $\bar{t}_n := \sum_{\ell=0}^{n} |b_\ell|$. By assumption, both $(\bar{s}_n)_{n\in\mathbb{N}}$ and $(\bar{t}_n)_{n\in\mathbb{N}}$ are convergent. With $A := \lim_{n\to\infty} \bar{s}_n$ and $B := \lim_{n\to\infty} \bar{t}_n$ it follows that $(\bar{s}_n \bar{t}_n)_{n\in\mathbb{N}}$ converges to $A \cdot B$. Let $\varepsilon > 0$ be arbitrary. Since $(\bar{s}_n \bar{t}_n)_{n\in\mathbb{N}}$, it is a Cauchy sequence, such that there exists $n_0 \in \mathbb{N}$ with

$$|\bar{s}_n \bar{t}_n - \bar{s}_m \bar{t}_m| < \varepsilon \qquad \forall m, n \geq n_0.$$

Put $n_1 := 2n_0$. Then combining the previous inequality with the above estimate yields

$$|s_n \cdot t_n - u_n| \leq |\bar{s}_n \bar{t}_n - \bar{s}_{n_0} \bar{t}_{n_0}| < \varepsilon \qquad \forall n \geq n_1.$$

Thus, $\lim_{n\to\infty}(s_n \cdot t_n) = \lim_{n\to\infty} u_n$. In particular, the limit exists, since $(s_n)_{n\in\mathbb{N}}$ and $(t_n)_{n\in\mathbb{N}}$ converge and it follows that

$$\sum_{n=0}^{\infty} c_n = \lim_{n\to\infty} u_n = \Big(\lim_{n\to\infty} s_n\Big) \cdot \Big(\lim_{n\to\infty} t_n\Big) = \Big(\sum_{n=0}^{\infty} a_n\Big) \cdot \Big(\sum_{n=0}^{\infty} b_n\Big).$$

The absolute convergence of $\sum_{n=0}^{\infty} c_n$ now follows by applying this result to the series $\sum_{n=0}^{\infty} |a_n|$, $\sum_{n=0}^{\infty} |b_n|$ and $\sum_{n=0}^{\infty} |c_n|$. □

Definition 2.85. In the situation of Theorem 2.84, the series $\sum_{n=0}^{\infty} c_n$ is called *the Cauchy product of* $\sum_{n=0}^{\infty} a_n$ *and* $\sum_{n=0}^{\infty} b_n$.

2.5 Power series and the exponential series

In this final section of Chapter 2, we want to study a particular type of series in detail that will be of great importance in the following chapter.

Definition 2.86. A series of the form

$$\sum_{n=0}^{\infty} a_n z^n, \qquad z \in \mathbb{C},$$

where $a_n \in \mathbb{C}$ for all $n \in \mathbb{N}_0$, is called a *power series*. a_n is called *the n-th coefficient* of the power series.

Example 2.87. The geometric series $\sum_{n=0}^{\infty} z^n$ is a power series. Its coefficients are given by $a_n = 1$ for all $n \in \mathbb{N}$.

In Chapter 3 we will consider power series as functions in the variable z and study their behaviour under varying z. In this section we want to investigate how to determine for which choices of z a power series converges. Obviously for $z = 0$ it holds that $\sum_{n=0}^{\infty} a_n z^n = a_0$, hence every power series converges in 0. For the example of the geometric series, we have seen that it converges if $|z| < 1$ and diverges if $|z| > 1$. . It turns out that the general convergence behaviour of power series is somewhat similar to this example.

Theorem 2.88. *Let* $\sum_{n=0}^{\infty} a_n z^n$ *be a power series and let*

$$r := \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

a) If $r \in (0, +\infty)$, *then* $\sum_{n=0}^{\infty} a_n z^n$ *converges absolutely if* $|z| < \frac{1}{r}$ *and diverges if* $|z| > \frac{1}{r}$.

b) If $r = 0$, *then* $\sum_{n=0}^{\infty} a_n z^n$ *converges absolutely for all* $z \in \mathbb{C}$.

c) If $r = +\infty$, *then* $\sum_{n=0}^{\infty} a_n z^n$ *converges only for* $z = 0$.

Definition 2.89. In the situation of Theorem 2.88, we call the number

$$\rho := \begin{cases} \frac{1}{r} & \text{if } r \in (0, +\infty), \\ 0 & \text{if } r = +\infty, \\ +\infty & \text{if } r = 0, \end{cases}$$

the *radius of convergence* of $\sum_{n=0}^{\infty} a_n z^n$.

Proof of Theorem 2.88. a) We want to apply the root test (Theorem 2.80) to the series $\sum_{n=0}^{\infty} a_n z^n$. Since

$$\limsup_{n\to\infty} \sqrt[n]{|a_n z^n|} = \limsup_{n\to\infty} \sqrt[n]{|a_n| \cdot |z|^n} = (\limsup_{n\to\infty} \sqrt[n]{|a_n|})|z| = r \cdot |z|.$$

The claim now follows immediately from Theorem 2.80.

b) If $r = 0$, then the computation from part a) shows that $\limsup_{n\to\infty} \sqrt[n]{|a_n z^n|} = 0$ for all $z \in \mathbb{C}$, so the claim again follows from Theorem 2.80.

c) If $(\sqrt[n]{|a_n|})_{n\in\mathbb{N}}$ is unbounded, then $(\sqrt{|a_n|\cdot|z|^n})_{n\in\mathbb{N}}$ is unbounded for each $z\in\mathbb{C}$ with $z\neq 0$. Hence, $\limsup_{n\to\infty}\sqrt[n]{|a_n z^n|} = +\infty$ for all $z\in\mathbb{C}\setminus\{0\}$ and the claim follows again from Theorem 2.80.

□

Given $\rho\in(0,+\infty)$ and viewing $\mathbb{C}$ as a two-dimensional plane, the set $\{z\in\mathbb{C}\mid |z|<\rho\}$ is the interior of a circle of radius ρ around the origin. Hence, Theorem 2.88 says that in the case of finite radius of convergence ρ, the power series converges for all z inside the circle of radius ρ and diverges for all z outside the circle.

Remark 2.90. Given a power series $\sum_{n=0}^{\infty} a_n z^n$ with finite positive radius of convergence ρ, it is not possible to make a general statement about convergence or divergence at those z with $|z|=\rho$. See the following example for the investigation of a special case.

Example 2.91. (1) Consider the power series $\displaystyle\sum_{n=1}^{\infty}\frac{z^n}{n}$. Since

$$\lim_{n\to\infty}\sqrt[n]{\frac{1}{n}} = \frac{1}{\lim_{n\to\infty}\sqrt[n]{n}} = 1,$$

its radius of convergence is 1. Thus, $\sum_{n=1}^{\infty}\frac{z^n}{n}$ converges for all $z\in\mathbb{C}$ with $|z|<1$. However, in the case $|z|=1$, the convergence behaviour becomes more delicate. We have seen in Example 2.67 that the series diverges for $z=1$, but we have also seen in Example 2.73 that the series converges for $z=-1$.

(2) Consider the series $\displaystyle\sum_{n=0}^{\infty} z^{n!} = z+z+z^2+z^6+z^{24}+\dots$.

This series is a power series with coefficients

$$a_k = \begin{cases} 1 & \text{if } k=n! \text{ for some } n\in\mathbb{N},\\ 0 & \text{else.}\end{cases}$$

Since $\lim_{n\to\infty}\sqrt[n!]{a_{n!}} = 1$ and since there are no subsequences converging against a bigger limit or diverging to infinity, the radius of convergence of $\sum_{n=0}^{\infty} z^{n!}$ is 1.

The special kind of behaviour of power series may sometimes be used to derive the convergence of series from the convergence of another series that does not look all to related to it upon first sight, for which one might use the following observation.

Corollary 2.92. *Assume that the power series $\sum_{n=0}^{\infty} a_n z^n$ converges in $z_0\in\mathbb{C}$. Then it converges absolutely for all $z\in\mathbb{C}$ with $|z|<|z_0|$.*

Proof. Let ρ be the radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$. Since $\sum_{n=0}^{\infty} a_n z_0^n$ converges, it follows from part of Theorem 2.88.a) that $\rho\geq|z_0|$. The claim then follows from Theorem 2.88. □

The surprising fact behind Corollary 2.92 is that convergence in one particular point z_0 implies convergence inside a whole circle.

Example 2.93. Let $\sum_{n=0}^{\infty} a_n$ be a convergent complex series. Then $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely for all $z\in\mathbb{C}$ with $|z|<1$. This follows from Corollary 2.92 with $z_0=1$, since $\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} a_n 1^n$.

Under certain conditions, there is a more convenient way of computing the radius of convergence of a power series.

Theorem 2.94 (Euler)**.** *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with $a_n \neq 0$ for all $n \in \mathbb{N}$. If the sequence $(\frac{|a_{n+1}|}{|a_n|})_{n\in\mathbb{N}}$ converges, then the radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$ is given by*

$$\rho = \lim_{n\to\infty} \frac{|a_n|}{|a_{n+1}|}.$$

Proof. (This proof is not discussed in the lecture videos.)
We apply the ratio test (Theorem 2.82) to the power series. Put $q := \lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$. Then

$$\limsup_{n\to\infty} \frac{|a_{n+1}z^{n+1}|}{|a_n z^n|} = \Big(\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}\Big) \cdot |z| = q \cdot |z|.$$

Hence, if $|z| < \frac{1}{q}$, then $\sum_{n=1}^{\infty} a_n z^n$ converges absolutely by the quotient test. Assume now that $|z| > \frac{1}{q}$, i.e. $|q| > 1$. Since $(\frac{|a_{n+1}|}{|a_n|})_{n\in\mathbb{N}}$ converges, this implies that $\frac{|a_{n+1}|}{|a_n|}|z| = \frac{|a_{n+1}z^{n+1}|}{|a_n z^n|} > 1$ for almost every $n \in \mathbb{N}$. Thus, by Theorem 2.82, the series $\sum_{n=1}^{\infty} a_n z^n$ diverges if $|z| > \frac{1}{q}$. It follows that (using the convention that $\frac{1}{0} = +\infty$ the radius of convergence is given by

$$\rho = \frac{1}{q} = \frac{1}{\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}} = \lim_{n\to\infty} \frac{|a_n|}{|a_{n+1}|}$$

by Theorems 2.14 and 2.43.b). □

The following theorem summarizes some results from the previous section for the special case of power series.

Theorem 2.95. *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence ρ_1 and $\sum_{n=0}^{\infty} b_n z^n$ be a power series with radius of convergence ρ_2. Put $\rho := \min\{\rho_1, \rho_2\}$. Then:*

a) $\sum_{n=0}^{\infty}(a_n + b_n)z^n$ converges for all $z \in \mathbb{C}$ with $|z| < \rho$ and for any such z it holds that

$$\sum_{n=0}^{\infty}(a_n + b_n)z^n = \sum_{n=0}^{\infty} a_n z^n + \sum_{n=0}^{\infty} b_n z^n.$$

b) For all $z \in \mathbb{C}$ with $|z| < \rho$ it holds that

$$\sum_{n=0}^{\infty}\Big(\sum_{k=0}^{n} a_k b_{n-k}\Big)z^n = \Big(\sum_{n=0}^{\infty} a_n z^n\Big) \cdot \Big(\sum_{n=0}^{\infty} b_n z^n\Big).$$

Proof. a) This follows directly from Theorem 2.70.a).

b) This follows directly from Theorem 2.84.

□

In the remainder of this section, we want to introduce a very important power series and discuss some of its most important properties. This series will be studied in greater detail in the next chapter.

Definition 2.96. The *exponential series* is the power series

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Theorem 2.97. $\exp(z)$ *is absolutely convergent for all* $z \in \mathbb{C}$.

Proof. Denote the coefficients of $\exp(z)$ by $a_n = \frac{1}{n!}$. Then

$$\frac{|a_{n+1}|}{|a_n|} = \frac{n!}{(n+1)!} = \frac{1}{n+1} \qquad \forall n \in \mathbb{N},$$

hence $\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = 0$. Thus, by Theorem 2.94 and Theorem 2.43.b), the radius of convergence of $\exp(z)$ is

$$\rho = \lim_{n\to\infty} \frac{|a_n|}{|a_{n+1}|} = +\infty,$$

which shows the claim. □

At this point, we are now able to define the number e, which is probably known to most of you from high school, where one encounters e^x as a differentiable function whose derivative is given by e^x itself. (We will define and study this function later on.)

Definition 2.98. The number

$$e := \exp(1) = \sum_{n=0}^{\infty} \frac{1}{n!}$$

is known as *Euler's number*.

Since

$$e = \frac{1}{0!} + \frac{1}{1!} + \sum_{n=2}^{\infty} \frac{1}{n!} = 2 + \sum_{n=2}^{\infty} \frac{1}{n!},$$

we can immediately see that $e > 2$.

The following theorem summarizes some of the most important properties of the exponential series.

Theorem 2.99. *a)* $\exp(0) = 1$.

b) For all $z, w \in \mathbb{C}$ *it holds that*

$$\exp(z+w) = \exp(z) \cdot \exp(w).$$

c) For all $z \in \mathbb{C}$ *it holds that*

$$\exp(-z) = \exp(z)^{-1}.$$

d) $\exp(z) \neq 0$ *for all* $z \in \mathbb{C}$ *and* $\exp(x) \in (0, +\infty)$ *for all* $x \in \mathbb{R}$.

e) For each $q \in \mathbb{Q}$ *it holds that* $\exp(q) = e^q$.

f) For all $z \in \mathbb{C}$ *it holds that* $\exp(\bar{z}) = \overline{\exp(z)}$.

g) For each $t \in \mathbb{R}$ *it holds that* $|\exp(it)| = 1$.

Proof. a) This is obvious.

b) Using Theorem 2.84 and the binomial theorem we compute that

$$\begin{aligned}\exp(z)\cdot\exp(w) &= \Big(\sum_{n=0}^{\infty}\frac{z^n}{n!}\Big)\cdot\Big(\sum_{n=0}^{\infty}\frac{w^n}{n!}\Big) = \sum_{n=0}^{\infty}\Big(\sum_{k=0}^{n}\frac{z^k}{k!}\frac{w^{n-k}}{(n-k)!}\Big)\\ &= \sum_{n=0}^{\infty}\Big(\sum_{k=0}^{n}\frac{1}{n!}\binom{n}{k}z^k w^{n-k}\Big) = \sum_{n=0}^{\infty}\frac{1}{n!}\Big(\sum_{k=0}^{n}\binom{n}{k}z^k w^{n-k}\Big)\\ &= \sum_{n=0}^{\infty}\frac{(z+w)^n}{n!} = \exp(z+w).\end{aligned}$$

c) We derive from a) and b) that for every $z\in\mathbb{C}$ it holds that

$$\exp(z)\cdot\exp(-z) = \exp(z-z) = \exp(0) = 1.$$

The claim immediately follows.

d) The first part follows directly from c). For the second part,it follows from the definition of exp that $\exp(x)\in(0,+\infty)$ if $x\geq 0$, since each summand of the series is positive. For $x<0$ we obtain by c) that $\exp(-x) = (\exp(x))^{-1}\in(0,+\infty)$.

e) We first want to show by induction that $\exp(n)\overset{!}{=} e^n$ for $n\in\mathbb{N}$.

The base case $n=1$ is true by definition of e. For the induction step, we assume that $\exp(n)=e^n$ for some $n\in\mathbb{N}$ and derive from this hypothesis that

$$\exp(n+1)\overset{\text{b)}}{=}\exp(n)\cdot\exp(1) = e^n\cdot e = e^{n+1},$$

which completes the induction step. Hence, $\exp(n)=e^n$ for each $n\in\mathbb{N}$. By c), this implies

$$\exp(-n) = (\exp(n))^{-1} = (e^n)^{-1} = e^{-n},$$

which proves that $\exp(m)=e^m$ for every $m\in\mathbb{Z}$.

Let now $n\in\mathbb{N}$ be arbitrary. We compute that

$$e = \exp(1) = \exp\Big(\frac{n}{n}\Big)\overset{(*)}{=}\Big(\exp\Big(\frac{1}{n}\Big)\Big)^n,$$

where $(*)$ is shown by induction in analogy with b). From d) and by taking n-th roots, this shows that

$$\exp\Big(\frac{1}{n}\Big) = e^{\frac{1}{n}}\quad\forall n\in\mathbb{N}.$$

Finally, for $q=\frac{m}{n}\in\mathbb{Q}$, we derive that

$$\exp(q) = \exp\Big(\frac{m}{n}\Big) = \Big(\exp\Big(\frac{1}{n}\Big)\Big)^m = \Big(e^{\frac{1}{n}}\Big)^m = e^{\frac{m}{n}} = e^q.$$

f) Let $s_n(z)$ be the n-th partial sum of $\exp(z)$, $z\in\mathbb{C}$. Then by the properties of complex conjugation

$$s_n(\bar{z}) = 1+\sum_{k=1}^{n}\frac{\bar{z}^k}{k!} = 1+\sum_{k=1}^{n}\frac{\overline{z^k}}{k!} = \overline{\sum_{k=0}^{n}\frac{z^k}{k!}} = \overline{s_n(z)}.$$

Hence,

$$\exp(\bar{z}) = \lim_{n\to\infty} s_n(\bar{z}) = \lim_{n\to\infty} \overline{s_n(z)} = \overline{\exp(z)},$$

which we wanted to show.[4]

g) By c) it holds for all $t \in \mathbb{R}$ that

$$1 = \exp(it) \cdot \exp(-it) \overset{\text{f)}}{=} \exp(it) \cdot \overline{\exp(it)} = |\exp(it)|^2 .$$

The claim immediately follows.

□

[4]Here, we have actually cheated a little. We have not established yet that if $\lim_{n\to\infty} z_n = z$, then $\lim_{n\to\infty} \bar{z}_n = \bar{z}$. However, this easiy follows from the properties of limits that we have discussed.

Chapter 3

Functions and continuity

3.1 Definition and first properties of continuity

In this section we are going to introduce the next important fundamental concept of analysis: continuity. In high school, continuity is often studied in a heuristic way, e.g. by saying that a function is continuous if its graph can be drawn without having to lift the tip of the pen. Roughly, a function $f : D \to \mathbb{C}$, where $D \subset \mathbb{C}$ is continuous in a point a, if for all points sufficiently close to a, the values of the function become very close to one another. Similar as in the case of convergence, we will formalize this by using ε-"neighborhoods" of points.

Definition 3.1. Let $D \subset \mathbb{C}$.

a) A *real function on D* is a map $f : D \to \mathbb{R}$. A *complex function on D* is a map $f : D \to \mathbb{C}$. Since $\mathbb{R} \subset \mathbb{C}$, every real function on D is a complex function on D.

b) Let $a \in D$. A complex function $f : D \to \mathbb{C}$ is *continuous in a* if

$$\forall \varepsilon > 0 \; \exists \delta > 0 : \quad |f(x) - f(a)| < \varepsilon \text{ for each } x \in D \text{ with } |x - a| < \delta.$$

c) $f : D \to \mathbb{C}$ is called *continuous* if f is continuous in a for every $a \in D$.

See the lecture video to this section for a graphic illustration of this definition.

Example 3.2. (1) Every constant function $f : \mathbb{C} \to \mathbb{C}$, $f(z) = c$, where $c \in \mathbb{C}$, is continuous. This is trivial since $|f(x) - f(y)| = 0$ for all $x, y \in \mathbb{C}$.

(2) $f : \mathbb{C} \to \mathbb{C}$, $f(z) = z$, is continuous.
Given $a \in \mathbb{C}$ and an arbitrary $\varepsilon > 0$ it follows for $\delta = \varepsilon$ that

$$|f(x) - f(a)| = |x - a| < \varepsilon \text{ for each } x \in D \text{ with } |x - a| < \delta.$$

(3) $f : \mathbb{C} \to \mathbb{C}$, $f(z) = z^2$, is continuous.
Let $a \in \mathbb{C}$. By the triangle inequality,

$$|x + a| = |x - a + 2a| \leq |x - a| + 2|a|$$

for all $x \in \mathbb{C}$. Hence, we derive that

$$|f(x) - f(a)| = |x^2 - a^2| = |x + a| \cdot |x - a| \leq |x - a|^2 + 2|a| \cdot |x - a|.$$

For $\varepsilon > 0$ put $\delta := \min\{1, \frac{\varepsilon}{2|a|+1}\}$. Then for all $x \in \mathbb{C}$ with $|x-a| < \delta$ we derive from $\delta \leq 1$ and the above that

$$|f(x) - f(a)| \leq |x-a| + 2|a| \cdot |x-a| < (2|a|+1)\delta \leq (2|a|+1)\frac{\varepsilon}{2|a|+1} = \varepsilon.$$

Thus, we have shown that f is continuous in a.

The definition of continuity is sometimes hard to handle, since it might be highly non-trivial to explicitly find a suitable $\delta > 0$ to each $\varepsilon > 0$. The following theorem not only provides a sometimes more practial criterion for continuity, but also connects continuity with the convergence of sequences.

Theorem 3.3 (Sequential continuity)**.** *Let $D \subset \mathbb{C}$ and $a \in D$. A complex function $f : D \to \mathbb{C}$ is continuous in a if and only if the following holds: For every convergent sequence $(x_n)_{n\in\mathbb{N}}$ in D with $\lim\limits_{n\to\infty} x_n = a$ the sequence $(f(x_n))_{n\in\mathbb{N}}$ converges with $\lim\limits_{n\to\infty} f(x_n) = f(a)$.*

Proof. "⇒": Let f be continuous in a and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in D with $\lim_{n\to\infty} x_n = a$. By definition of convergence, we need to show that

$$\forall \varepsilon > 0\ \exists n_0 \in \mathbb{N} : \quad |f(x_n) - f(a)| \overset{!}{<} \varepsilon\ \ \forall n \geq n_0.$$

Let $\varepsilon > 0$. Since f is continuous in a there exists $\delta > 0$, such that:

$$|x-a| < \delta \quad \Rightarrow \quad |f(x) - f(a)| < \varepsilon.$$

Since $(x_n)_{n\in\mathbb{N}}$ converges to a, there exists $n_0 \in \mathbb{N}$ with $|x_n - a| < \delta$ for all $n \geq n_0$. Thus, by the previous implication, $|f(x_n) - f(a)| < \varepsilon$ for all $n \geq n_0$. Since $\varepsilon > 0$ was arbitrary, $(f(x_n))_{n\in\mathbb{N}}$ converges to $f(a)$.
"⇐": Assume that f satisfies the sequence condition in a, but is not continuous in a. Then there exists $\varepsilon_0 > 0$, such that for each $\delta > 0$ there exists an $x \in D$ with $|x-a| < \delta$, but $|f(x) - f(a)| \geq \varepsilon_0$. In particular, for each $n \in \mathbb{N}$ there exists an $x_n \in \mathbb{N}$ with $|x_n - a| < \frac{1}{n}$, but $|f(x_n) - f(a)| \geq \varepsilon_0$. By construction, the resulting sequence $(x_n)_{n\in NN}$ satisfies $\lim_{n\to\infty} x_n = a$, but since $\lim_{n\to\infty} |f(x_n) - f(a)| \geq \varepsilon_0 \neq 0$ by Corollary 2.27, the sequence $(f(x_n))_{n\in\mathbb{N}}$ can not converge to $f(a)$. This contradicts the assumption, hence f is continuous in a. □

Example 3.4. The *Heaviside function*[1]

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0, \end{cases}$$

is *not* continuous in 0. Given a real sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n < 0$ for all $n \in \mathbb{N}$ and with $\lim_{n\to\infty} x_n = 0$, it holds that

$$\lim_{n\to\infty} f(x_n) = 0 \neq 1 = f(0).$$

Hence, f is not continuous in 0 by Theorem 3.3.

In the same way as we did for convergence in the previous chapter, we are next going to study the behaviour of continuity with respect to algebraic operations on $\mathbb{R}$ and $\mathbb{C}$.

[1] Fun fact: This function is not named that way since it has a "heavy side" on $[0, +\infty)$, but it is indeed named after the mathematician and physicist Oliver Heaviside (1850-1925).

Theorem 3.5. *Let $D \subset \mathbb{C}$, $a \in D$ and let $f, g : D \to \mathbb{C}$ both be continuous in a. Then:*

a) $f + g : D \to \mathbb{C}$, $(f+g)(x) = f(x) + g(x)$, *is continuous in a.*

b) $f \cdot g : D \to \mathbb{C}$, $(f \cdot g)(x) = f(x) \cdot g(x)$, *is continuous in a.*

c) If $f(x) \neq 0$ *for all* $x \in D$, *then* $\frac{1}{f} : D \to \mathbb{C}$, $\frac{1}{f}(x) = \frac{1}{f(x)}$, *is continuous in a.*

d) $\mathrm{Re}(f) : D \to \mathbb{R}$, $(\mathrm{Re}(f))(x) := \mathrm{Re}(f(x))$, *and* $\mathrm{Im}(f) : D \to \mathbb{R}$, $(\mathrm{Im}(f))(x) := \mathrm{Im}(f(x))$, *are continuous in a.*

Proof. a) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in D converging to a. Then by Theorem 3.3 and Theorem 2.7.a) about sums of sequences:

$$\lim_{n\to\infty}(f+g)(x_n) = \lim_{n\to\infty}(f(x_n)+g(x_n)) = \lim_{n\to\infty} f(x_n) + \lim_{n\to\infty} g(x_n) = f(a)+g(a) = (f+g)(a).$$

Hence, $f + g$ is continuous in a by Theorem 3.3.

b) This follows in analogy with part a) from Theorem 2.12.b) about products of sequences.

c) This follows in analogy with part a) using Theorem 2.14.

d) This follows in analogy with part a) using Theorem 2.20.

□

Corollary 3.6. *Every complex polynomial, i.e every complex function of the form $f : D \to \mathbb{C}$, $f(x) = \sum_{k=0}^{n} a_k z^k$, where $D \subset \mathbb{C}$, $n \in \mathbb{N}$, $a_0, a_1, \ldots, a_n \in \mathbb{C}$, is continuous.*

Proof. Apply Theorem 3.5.a) and b) to the functions from Example 3.2.(1) and (2). □

Theorem 3.5 will be used over and over in the course of this lecture. In the same way, we will often use the following theorem about the compatibility of continuity and compositions of maps.

Theorem 3.7. *Let $D_1, D_2 \subset \mathbb{C}$, let $a \in D_1$ and let $f : D_1 \to D_2$ and $g : D_2 \to \mathbb{C}$ be functions. If f is continuous in a and g is continuous in $f(a)$, then $g \circ f : D_1 \to \mathbb{C}$ is continuous in a.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in D_1 converging to a. Since f is continuous, it follows that $(f(x_n))_{n\in\mathbb{N}}$ converges to $f(a)$. Since g is continuous, it follows that

$$\lim_{n\to\infty}(g\circ f)(x_n) = \lim_{n\to\infty} g(f(x_n)) = g(f(a)) = (g\circ f)(a).$$

Hence, $g \circ f$ is continuous in a by Theorem 3.3. □

We next want to introduce a stricter notion of continuity that is often much easier to check than the original one. Nevertheless, it is stricter than the general definition as we shall see in Remark 3.11.

Definition 3.8. Let $D \subset \mathbb{C}$. A map $f : D \to \mathbb{C}$ is called *Lipschitz-continuous* if there exists $L \geq 0$, such that

$$|f(x) - f(y)| \leq L \cdot |x - y| \qquad \forall x, y \in D.$$

An L with this property is called a *Lipschitz constant for* f.

Example 3.9. (1) $f : \mathbb{C} \to \mathbb{R}$, $f(z) = |z|$, is Lipschitz-continuous with Lipschitz constant 1. One derives from the triangle inequality that $||x| - |y|| \leq |x - y|$ for all $x, y \in \mathbb{C}$, which reads as

$$|f(x) - f(y)| \leq 1 \cdot |x - y| \qquad x, y \in \mathbb{C}.$$

(2) The complex conjugation $f : \mathbb{C} \to \mathbb{C}$, $f(z) = \bar{z}$, is Lipschitz-continuous with Lipschitz constant 1. This easily follows from Theorem 1.73.b).

(3) Given $a, b \in \mathbb{C}$ let $f : \mathbb{C} \to \mathbb{C}$, $f(x) = ax + b$. Then for all $x, y \in \mathbb{C}$ it holds that

$$|f(x) - f(y)| = |ax + b - ay - b| = |a(x - y) = |a| \cdot |x - y|,$$

hence f is Lipschitz-continuous with Lipschitz constant $|a|$.

Theorem 3.10. *Let $D \subset \mathbb{C}$. Every Lipschitz-continuous complex function on D is continuous.*

Proof. Let $f : D \to \mathbb{C}$ and let $L > 0$, such that $|f(x) - f(y)| \leq L \cdot |x - y|$ for all $x, y \in D$, and let $a \in D$.
Given $\varepsilon > 0$, we put $\delta := \frac{\varepsilon}{L}$. Then for each $x \in D$ with $|x - a| < \delta$ it holds that

$$|f(x) - f(a)| \leq L \cdot |x - a| < L \cdot \frac{\varepsilon}{L} = \varepsilon.$$

Hence, f is continuous in a. □

Remark 3.11. The converse of Theorem 3.10 is false in that there are continuous complex functions that are not Lipschitz-continuous. As an example consider the function

$$f : [0, +\infty) \to \mathbb{R}, \qquad f(x) = \sqrt{x}.$$

We will see in the next section that f is indeed continuous. However, assume that f was Lipschitz-continuous with Lipschitz constant $L > 0$. Then, it would follow that

$$|f(\tfrac{1}{4L^2}) - f(0)| \leq L \cdot |\tfrac{1}{4L^2} - 0| \quad \Leftrightarrow \quad \frac{1}{2L} \leq \frac{1}{4L} \quad \Leftrightarrow \quad \frac{1}{2} \leq \frac{1}{4},$$

which is obviously false. Hence, f is not Lipschitz-continuous.

Corollary 3.12. *Let $D \subset \mathbb{C}$ and $a \in D$.*

a) If $f : D \to \mathbb{C}$ is continuous in a, then $|f| : D \to \mathbb{R}$, $|f|(x) := |f(x)|$, is continuous in a.

b) If $f, g : D \to \mathbb{R}$ are continuous, then $\max(f, g) : D \to \mathbb{R}$, $(\max(f, g))(x) := \max\{f(x), g(x)\}$, and $(\min(f, g)) : D \to \mathbb{R}$, $(\min(f, g))(x) := \min\{f(x), g(x)\}$, are continuous.

Proof. a) By Example 3.9 and Theorem 3.10, $g : \mathbb{C} \to \mathbb{R}$, $g(x) = |x|$ is continuous. The claim thus follows from Theorem 3.7.

b) One checks without difficulties that all $x, y \in \mathbb{R}$ satisfy

$$\max\{x, y\} = \frac{1}{2}(x + y + |x - y|), \qquad \min\{x, y\} = \frac{1}{2}(x + y - |x - y|).$$

This implies

$$\max(f, g) = \frac{1}{2}(f + g + |f - g|), \qquad \min(f, g) = \frac{1}{2}(f + g - |f - g|),$$

so the claim follows from part a) and Theorem 3.5.

□

An important class of Lipschitz-continuous functions are power series $\sum_{n=0}^{\infty} a_n z^n$, seen as functions in z. For $r > 0$ we let $D_r := \{z \in \mathbb{C} \mid |z| \leq r\}$. We further put $D_0 := \{0\}$ and $D_{+\infty} := \mathbb{C}$.

Theorem 3.13. *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series of radius of convergence ρ. Then for each $r < \rho$, the map $f : D_r \to \mathbb{C}$, $f(z) = \sum_{n=0}^{\infty} a_n z^n$, is Lipschitz-continuous.*

Proof. Let $s_n(z) := \sum_{k=0}^{n} a_k z^k$ be the n-th partial sum of $f(z)$. As a polynomial, $s_n : D_r \to \mathbb{C}$ is continuous for each $n \in \mathbb{N}$. Since $f(z) = \lim_{n\to\infty} s_n(z)$ for each $n \in \mathbb{N}$, it holds that

$$f(z) - f(w) = \lim_{n\to\infty} (s_n(z) - s_n(w)) \qquad \forall z, w \in D_r.$$

By Corollary 3.12.a), this implies

$$|f(z) - f(w)| = \lim_{n\to\infty} |s_n(z) - s_n(w)|. \tag{3.1}$$

By the triangle inequality,

$$|s_n(z) - s_n(w)| = \Big| \sum_{k=1}^{n} a_k (z^k - w^k) \Big| \leq \sum_{k=1}^{n} |a_k| \cdot |z^k - w^k|$$

for every $n \in \mathbb{N}$. For each $k \in \mathbb{N}$ one checks that

$$z^k - w^k = (z - w) \cdot \sum_{j=0}^{k-1} z^j w^{k-1-j}.$$

Thus,

$$\left|z^k - w^k\right| \leq |z - w| \sum_{j=0}^{k-1} |z|^j |w|^{k-1-j} \leq |z - w| \sum_{j=0}^{k-1} r^j r^{k-1-j} = |z - w| \cdot k r^{k-1}.$$

Inserting this into the previous computation, we obtain

$$|s_n(z) - s_n(w)| \leq \sum_{k=1}^{n} |a_k| |z - w| k r^{k-1} = \sum_{k=1}^{n} |a_k| k \rho^{k-1} \cdot |z - w|.$$

By (3.1), this implies

$$|f(z) - f(w)| \leq \lim_{n\to\infty} \sum_{k=1}^{n} |a_k| k r^{k-1} \cdot |z - w| = \sum_{n=0}^{\infty} |a_{n+1}| (n+1) r^n \cdot |z - w|.$$

It only remains to show that the series $\sum_{n=1}^{\infty} |a_n| n r^{n-1}$ converges to some $L > 0$, which is then a Lipschitz constant for f. We compute that[2]

$$\limsup_{n\to\infty} \sqrt[n]{|a_n| n} = \limsup_{n\to\infty} \sqrt[n]{|a_n|} \sqrt[n]{n} = \limsup_{n\to\infty} \sqrt[n]{|a_n|} \cdot \lim_{n\to\infty} \sqrt[n]{n} = \frac{1}{\rho} \cdot 1 = \frac{1}{\rho}.$$

Thus, the radius of convergence of $\sum_{n=1}^{\infty} a_n n x^n$ is given by ρ as well. Since $r < \rho$, this shows the convegence of $\sum_{n=1}^{\infty} |a_n| n r^n$, thereby showing the claim. □

[2]Here we use the fact that for a real sequences $(x_n)_{n\in\mathbb{N}}$ and a convergent real sequence $(y_n)_{n\in\mathbb{N}}$ it holds that

$$\limsup_{n\to\infty} (x_n \cdot y_n) = \limsup_{n\to\infty} x_n \cdot \lim_{n\to\infty} y_n.$$

We leave the proof of this simple fact to the reader.

3.2 Continuous real functions on intervals

In this section, we want to focus on one of the most common cases appearing in practical applications, real-valued functions on intervals. We begin with an important theorem that has many interesting applications.

Theorem 3.14 (Intermediate value theorem)**.** *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous. Then for every c between $f(a)$ and $f(b)$ there exists $\xi \in [a, b]$ with $f(\xi) = c$.*

Proof. We assume w.l.o.g.[3] that $f(a) \leq f(b)$. (Otherwise, replace f by $-f$.)
Let $c \in [f(a), f(b)]$. If $c \in \{f(a), f(b)\}$, then the claim is obvious, so assume that $f(a) < c < f(b)$ and consider the set

$$M := \{x \in [a, b] \mid f(x) \leq c\}.$$

M is non-empty since $a \in M$. Since b is an upper bound, M is bounded from above. Thus, by construction of $\mathbb{R}$, M possesses a supremum $\xi := \sup M$. Since b is an upper bound for M, $\xi \leq b$. Since $a \in M$, $a \leq \xi$, thus $\xi \in [a, b]$.
By definition, ξ is the smallest upper bound of M, hence for every $n \in \mathbb{N}$ there exists an $x_n \in M$ with $\xi - \frac{1}{n} \leq x_n \leq \xi$. If we choose such a sequence $(x_n)_{n\in\mathbb{N}}$ it follows from the squeeze theorem that $\lim_{n\to\infty} x_n = \xi$. Since f is continuous, $f(\xi) = \lim_{n\to\infty} f(x_n)$ and since $f(x_n) \leq c$ for each n by construction, it follows that $f(\xi) \leq c$.
To show the opposite inequality, we first note that since $c \neq f(b)$, it holds that $\xi < b$. Let $(y_n)_{n\in\mathbb{N}}$ be a convergent sequence with $y_n \in (\xi, b]$ for each $n \in \mathbb{N}$ and with $\lim_{n\to\infty} y_n = \xi$. Since $y_n > \xi$, it follows that $y_n \notin M$ and thus $f(y_n) > c$ for each $n \in \mathbb{N}$. Since f is continuous, this implies $f(\xi) = \lim_{n\to\infty} f(y_n) \geq c$. Thus, we have shown that $f(\xi) = c$. □

Remark 3.15. The continuity of f is essential in the intermediate value theorem. For example, the Heaviside function from Example 3.4 takes the values 0 and 1, but no element of $(0, 1)$ is among its values.

Corollary 3.16. *Let $P : \mathbb{R} \to \mathbb{R}$ be a real polynomial of degree n, i.e. $P(x) = \sum_{k=0}^{n} a_k x^k$, where $n \in \mathbb{N}$, $a_0, a_1, \ldots, a_n \in \mathbb{R}$ with $a_n \neq 0$. If n is odd, then there exists $x_0 \in \mathbb{R}$ with $P(x_0) = 0$.*

Proof. We assume w.l.o.g. that $a_n > 0$. (Otherwise consider $-P$ instead.) For every $m \in \mathbb{N}$ we compute that

$$P(m) = \sum_{k=0}^{n} a_k m^k = m^n \Big(a_n + \frac{a_{n-1}}{m} + \frac{a_{n-2}}{m^2} + \cdots + \frac{a_0}{m^n}\Big).$$

Since $\lim_{m\to\infty} m^n = +\infty$, it follows from Sheet 4, Exercise 1.b), that $\lim_{m\to\infty} P(m) = +\infty$. In particular, there exists $m_1 \in \mathbb{N}$ with $P(m_1) > 0$. We further compute that

$$\begin{aligned} P(-m) &= (-m)^n \Big(a_n - \frac{a_{n-1}}{m} + \frac{a_{n-2}}{m^2} - \cdots + (-1)^n \frac{a_0}{m^n}\Big) \\ &= -m^n \Big(a_n - \frac{a_{n-1}}{m} + \frac{a_{n-2}}{m^2} - \cdots + (-1)^n \frac{a_0}{m^n}\Big), \end{aligned}$$

since n is odd. This shows analogously that $\lim_{m\to\infty} P(-m) = -\infty$. Hence, there exists $m_2 \in \mathbb{N}$ with $P(-m_2) < 0$. It follows from the intermediate value theorem that there exists $x_0 \in [-m_2, m_1]$ with $P(x_0) = 0$. □

[3] w.l.o.g. stands for: **w**ithout **l**oss **o**f **g**enerality. This acronym is often used in mathematics, in situations where one can restrict the proof by proving a special case only and has the remtaining directly following from that case.

We can give another interpretation of the intermediate value theorem in terms of intervals. To do so, we first investigate another way of characterizing intervals.

Theorem 3.17. *Let $D \subset \mathbb{R}$. Then D is an interval if and only if it has the following property: For all $x_1, x_2 \in D$ with $x_1 \leq x_2$ it holds that $[x_1, x_2] \subset D$.*

Proof. "$\Rightarrow$": This follows immediately from the definition of intervals.
"$\Leftarrow$": We first consider the case that D is bounded, i.e. a subset of a bounded interval. If D has only one element, it is of the form $\{a\} = [a, a]$, hence an interval. Assume in the following that D has at least two elements.
Let $x \in (\inf D, \sup D)$. Then x is neither a lower nor an upper bound of D, hence there are $x_1, x_2 \in D$ with $x_1 < x < x_2$. But since $x \in [x_1, x_2]$, it follows by assumption that $x \in D$. Hence,

$$(\inf D, \sup D) \subset D \subset [\inf D, \sup D],$$

where the second inclusion is obvious.It follows from these inclusions that D is given by one of the intervals

$$[\inf D, \sup D], \qquad (\inf D, \sup D), \qquad [\inf D, \sup D), \qquad (\inf D, \sup D],$$

hence D is an interval.
In the case that D is bounded from below, but not from above, the analogous argument shows that $(\inf D, +\infty) \subset D \subset [\inf D, +\infty)$, such that $D \in \{(\inf D, +\infty), [\inf D, +\infty)\}$.
The case that D is bounded from above, but not from below is treated analogously, while if D is neither bounded from above nor from below, it follows that $D = \mathbb{R} = (-\infty, +\infty)$. □

Corollary 3.18. *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous. Then $f(I)$ is an interval.*

Proof. Let $y_1, y_2 \in f(I)$ with $y_1 \leq y_2$. Then by the intermediate value theorem, every $y \in [y_1, y_2]$ lies in $f(I)$ as well, i.e. $[y_1, y_2] \subset f(I)$. Hence, $f(I)$ is an interval by Theorem 3.17. □

Remark 3.19. The image of a continuous $f : I \to \mathbb{R}$ does not have to be an interval of the same type as I. For example, the image of

$$f : (-1, 1) \to \mathbb{R}, \qquad f(x) = x^2,$$

is given by $[0, 1)$, which is not open in contrast to $(-1, 1)$. Another example is

$$g : [1, +\infty) \to \mathbb{R}, \qquad g(x) = \frac{1}{x}.$$

Here, the domain of g is unbounded, while $g([1, +\infty)) = (0, 1]$ is bounded.

Our next aim is to investigate monotonic functions and their properties. This concept basically transfers the notion of monotonicity from sequences to functions.

Definition 3.20. Let I be an interval and let $f : I \to \mathbb{R}$ be a real function.

a) f is *monotonically increasing* if for all $x, y \in I$ with $x \leq y$ it holds that $f(x) \leq f(y)$.

b) f is *monotonically decreasing* if for all $x, y \in I$ with $x \leq y$ it holds that $f(x) \geq f(y)$.

c) f is *strictly monotonically increasing* if for all $x, y \in I$ with $x < y$ it holds that $f(x) < f(y)$.

d) f is *strictly monotonically decreasing* if for all $x, y \in I$ with $x < y$ it holds that $f(x) > f(y)$.

f is called *monotonic* if it is monotonically increasing or monotonically decreasing and it is called *strictly monotonic* if it is strictly monotonically increasing or stricly monotonically decreasing.

One derives from the definition that strictly monotonic functions are injective. Using this observation and the intermediate value thoerem, we can at this point a second, but notably shorter, proof of Theorem 1.48 about the existence of roots.

Second proof of Theorem 1.48. Uniqueness: Consider the function $f : [0, +\infty) \to \mathbb{R}$, $f(x) = x^n - a$. From the computational rules of ordered fields one derives that f is strictly monotonically increasing on $[0, +\infty)$. In particular, f is injective, so there can be at most one solution. x_0 of $f(x_0) = 0$.
Existence: Since $a \geq 0$, it holds that $f(0) = -a \leq 0$. By Bernoulli's inequality it further holds that

$$f(1 + a) = (1 + a)^n - a \geq 1 + (n - 1)a \geq 1.$$

Since f is continuous, it follows from the intermediate value theorem that there exists $x_0 \in [0, 1 + a]$ with $f(x_0) = 0$, i.e. with $x_0^n = a$. □

For continuous functions, injectivity and strict monotonicity are more closely related.

Theorem 3.21. *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous and injective. Then f is strictly monotonic.*

Proof. Assume that f is continuous and injective, but neither strictly monotonically increasing not strictly monotonically decreasing. Then there are $a, b, c, d \in I$ with $a < b$ and $c < d$, such that $f(a) < f(b)$ and $f(c) > f(d)$. Consider the function

$$g : [0, 1] \to \mathbb{R}, \qquad g(t) = f(a + t(c - a)) - f(b + t(d - b)).$$

g is well-defined since $a + t(c - a), b + t(d - b) \in I$ for each $t \in [0, 1]$ by Theorem 3.17.g is a composition of continuous functions, thus itself continuous. We compute that

$$g(0) = f(a) - f(b) < 0, \qquad g(1) = f(c) - f(d) > 0.$$

By the intermediate value theorem, there exists $t_0 \in [0, 1]$ with

$$g(t_0) = 0 \quad \Leftrightarrow \quad f(a + t_0(c - a)) = f(b + t_0(d - b)).$$

Since f is injective, this implies

$$a + t_0(c - a) = b + t_0(d - b) \quad \Leftrightarrow \quad (1 - t_0)a + t_0 c = (1 - t_0)b + t_0 d.$$

But since $a < b$ and $c < d$, it follows that $(1 - t_0)a + t_0 c < (1 - t_0)b + t_0 d$, so we have achieved a contradiction. Thus, f is strictly monotonic. □

Remark 3.22. The conclusion of Theorem 3.21 is in general false if one does not assume continuity of f. For example, one observes that

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

is injective (even bijective), but not monotonic.

Theorem 3.23. *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous and strictly monotonic. Then $J := f(I)$ is an interval and the inverse function $f^{-1} : J \to I$ is continuous and strictly monotonic. Either both f and f^{-1} are strictly monotonically increasing or both are stricly monotonically decreasing.*

Proof. We assume w.l.o.g. that f is strictly monotonically *increasing*, the decreasing case is treated analogously. By Corollary 3.18, J is an interval. By monotonicity, f is injective, hence is bijective as a map $I \to J$. Hence, $f^{-1} : J \to I$ is well-defined and the definition of strict monotonicity directly implies that it is strictly monotonically increasing as well. Let $(y_n)_{n\in\mathbb{N}}$ be a convergent sequence in J with $\lim_{n\to\infty} y_n = y_0 \in J$. As a convergent sequence, $(y_n)_{n\in\mathbb{N}}$ is bounded, hence $A := \inf\{y_n \mid n \in \mathbb{N}\}$ and $B := \sup\{y_n \mid n \in \mathbb{N}\}$ are well-defined real numbers.

Claim: $A, B \in J$.

Proof of claim. If $A, B \in \{y_n \mid n \in \mathbb{N}\}$, then the claim is clear. Assume that $A \notin \{y_n \mid n \in \mathbb{N}\}$. By definition of the infimum, A is then an accumulation points of $(y_n)_{n\in\mathbb{N}}$. By Theorem 2.55 there is a subsequence $(y_{n_k})_{k\in\mathbb{N}}$ converging to A. But since $(y_n)_{n\in\mathbb{N}}$ itself converges, this implies by Theorem 2.19 that

$$A = \lim_{k\to\infty} y_{n_k} = \lim_{n\in\mathbb{N}} y_n = y_0 \in J.$$

Analogously, one shows that $B \in J$. ■

Since J is an interval, $[A, B] \subset J = f(I)$ by Theorem 3.17. Let $a, b \in I$ with $f(a) = A$ and $f(b) = B$. Since f is strictly monotonically increasing, it follows that $a < b$ and the intermediate value theorem yields $f([a, b]) = [A, B]$. Hence, $(f^{-1}(y_n))_{n\in\mathbb{N}}$ is a sequence in $[a, b]$, so it is particularly bounded, thus has a convergent subsequence $(f^{-1}(y_{n_k}))_{k\in\mathbb{N}}$ by the Bolzano-Weierstraß theorem (Theorem 2.48). Put $a_k := f^{-1}(y_{n_k})$ for all $k \in \mathbb{N}$ and put $a := \lim_{k\to\infty} a_k$. Since f is continuous, it follows that

$$f(a) = \lim_{k\to\infty} f(a_k) = \lim_{k\to\infty} y_{n_k} = \lim_{n\to\infty} y_n = y_0.$$

Since f is injective, it follows from $f(a) = y_0$ that $a = f^{-1}(y_0)$. Since this holds for *every* convergent subsequence of $(y_n)_{n\in\mathbb{N}}$, it follows that $f^{-1}(y_0)$ is the only accumulation point of $(f^{-1}(y_n))_{n\in\mathbb{N}}$. Thus, applying Corollary 2.62, we obtain

$$\limsup_{n\to\infty} f^{-1}(y_n) = \liminf_{n\to\infty} f^{-1}(y_n) = f^{-1}(y_0) \quad \Rightarrow \quad \lim_{n\to\infty} f^{-1}(y_n) = f^{-1}(y_0).$$

It follows from Theorem 3.3 that f^{-1} is continuous. □

Remark 3.24. If the domain of f is not an interval in the situation of Theorem 3.23, then its inverse is *not* necessarily continuous. Consider the following illuminating example:

$$f : (-\infty, 0] \cup (1, +\infty) \to \mathbb{R}, \qquad f(x) = \begin{cases} x & \text{if } x \leq 0, \\ x - 1 & \text{if } x > 1. \end{cases}$$

One checks that $f((-\infty, 0] \cup (1, +\infty) = \mathbb{R}$, that f is strictly monotonically increasing, hence injective, and that its inverse is given by

$$g : \mathbb{R} \to (-\infty, 0] \cup (1, +\infty), \qquad g(x) = \begin{cases} x & \text{if } x \leq 0, \\ x + 1 & \text{if } x > 0. \end{cases}$$

which is not continuous in 0.

Corollary 3.25. *For each $n \in \mathbb{N}$ with $n \geq 2$, the function $g : [0, +\infty) \to [0, +\infty)$, $g(x) = \sqrt[n]{x}$, is continuous.*

Proof. This follows directly from Theorem 3.23, since $g = f^{-1}$, where $f : [0, +\infty) \to [0, +\infty)$, $f(x) = x^n$. □

3.3 Continuous functions on compact subsets

In this section we will investigate the special role of so-called compact subsets of $\mathbb{R}$ and $\mathbb{C}$. In particular, we will show that every real function on a compact subset possesses a maximum and a minimum, which will justify the search for minima and maxima of functions that we will tackle later in this lecture.

Definition 3.26. Let $A \subset \mathbb{R}$ or $A \subset \mathbb{C}$.

a) A is *bounded* if there exists an $R > 0$, such that $|x| \leq R$ for all $x \in A$.

b) A is *closed* if for every convergent complex sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n \in A$ for every $n \in \mathbb{N}$ it holds that $\lim_{n\to\infty} a_n \in A$.

c) We call A *compact* if it is closed and bounded.

Example 3.27. (1) Every interval of the form $[a, b]$, where $a, b \in \mathbb{R}$, is compact.

(2) The interval $(0, 1]$ is bounded, but not closed: the sequence $a_n = \frac{1}{n}$ satisfies $a_n \in (0, 1]$ for all $n \in \mathbb{N}$, but $\lim_{n\to\infty} a_n = 0 \notin (0, 1]$.

(3) An interval of the form $[a, +\infty)$ is closed, but not bounded. Closedness holds since for every convergent sequence $(x_n)_{n\in\mathbb{N}}$ in $[a, +\infty)$ with $x_n \geq a$ for each $n \in \mathbb{N}$ it holds that $\lim_{n\to\infty} x_n \geq a$.

A more sophisticated example is discussed in the next theorem.

Theorem 3.28. *Given $z_0 \in \mathbb{C}$ and $r > 0$, the set*

$$D_r(z_0) := \{z \in \mathbb{C} \mid |z - z_0| \leq r\}$$

is compact. (Geometrically $D_r(z_0)$ denotes the interior and the boundary of a circle of radius r with center z_0.)

Proof. Let $R := r + |z_0|$. Then for each $z \in D_r(z_0)$ it holds by the triangle inequality that

$$|z| \leq |z - z_0| + |z_0| \leq r + |z_0| = R.$$

Hence, $D_r(z_0)$ is bounded.
Let $(x_n)_{n\in\mathbb{N}}$ be a convergent sequence in $D_r(z_0)$ and let $x_0 := \lim_{n\to\infty} x_n$. Then $(x_n - z_0)_{n\in\mathbb{N}}$ converges to $x_0 - z_0$ and by Example 3.9.(1) and Theorem 3.10, it follows that

$$\lim_{n\to\infty} |x_n - z_0| = |x_0 - z_0|.$$

Since $|x_n - z_0| \leq r$ for all $n \in \mathbb{N}$, this implies $|x_0 - z_0|$, i.e. that $x_0 \in D_r(x_0)$. Hence, $D_r(x_0)$ is closed, thus compact. □

Theorem 3.29. *Let $K \subset \mathbb{C}$ or $K \subset \mathbb{R}$ be compact and let $f : K \to \mathbb{C}$ be continuous. Then $f(K)$ is compact.*

Proof. Let $(y_n)_{n\in\mathbb{N}}$ be a convergent sequence in $f(K)$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in K with $f(x_n) = y_n$ for each $n \in \mathbb{N}$. Since K is bounded, $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ by the Bolzano-Weierstraß theorem. Let $x_0 := \lim_{k\to\infty} x_{n_k}$. Since K is closed, $x_0 \in K$. Then, by Theorem 2.19 and the continuity of f,

$$\lim_{n\to\infty} y_n = \lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(x_{n_k}) = f(x_0).$$

Hence, $\lim_{n\to\infty} y_n \in f(K)$, which shows that $f(K)$ is closed.
Suppose that $f(K)$ was not bounded. Then for each $n \in \mathbb{N}$ there exists $z_n \in f(K)$ with $|z_n| \geq n$. Let $(w_n)_{n\in\mathbb{N}}$ be a sequence in K with $f(w_n) = z_n$ for all $n \in \mathbb{N}$. Since K is bounded, $(w_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(w_{n_k})_{k\in\mathbb{N}}$ by the Bolzano-Weierstraß theorem, such that by continuity of f, $(f(w_{n_k}))_{k\in\mathbb{N}} = (z_{n_k})_{k\in\mathbb{N}}$ must be convergent. But this can not be the case, since $\lim_{k\to\infty} |z_{n_k}| = +\infty$ by construction, so $(z_{n_k})_{k\in\mathbb{N}}$ is unbounded. Hence, we have achieved a contradiction, which shows that $f(K)$ is bounded. Thus, $f(K)$ is compact. □

The previous theorem, which might not look all too spectacular upon first sight, has some important consequences for real functions.

Definition 3.30. Let $D \subset \mathbb{C}$ and $f : D \to \mathbb{R}$ be a real function.

a) $M \in \mathbb{R}$ is called *the maximum of f* and denoted by $\max f$, if $f(x) \leq M$ for each $x \in D$ and if there exists $x_0 \in D$ with $f(x_0) = M$. One says that *f attains its maximum in x_0.* and calls x_0 a *maximum point of f*.

b) $m \in \mathbb{R}$ is called *the minimum of f* and denoted by $\min f$, if $f(x) \geq m$ for each $x \in D$ and if there exists $x_0 \in M$ with $f(x_0) = m$. One says that *f attains its minimum in x_0* and calls x_0 a *minimum point of f*.

Remark 3.31. (1) Not every continuous function possesses a maximum or a minimum. For example, the function $f : (0, 1) \to \mathbb{R}$, $f(x) = x$ has neither of the two.

(2) One checks from the definition that $f : D \to \mathbb{R}$ possesses a maximum if and only if $\sup f(D) \in f(D)$, then $\max f = \max f(D)$. Analogously, f possesses a minimum if and only if $\inf f(D) \in f(D)$, then $\min f = \min f(D)$.

Theorem 3.32. *Let $K \subset \mathbb{C}$ be compact and let $f : K \to \mathbb{R}$ be continuous. Then f possesses a maximum and a minimum.*

Proof. By Remark 3.31, it suffices to show that $\sup f(K), \inf f(K) \in f(K)$. Put $A := f(K)$. By Theorem 3.29, A is a compact subset of $\mathbb{R}$. In particular, $\sup A, \inf A \in \mathbb{R}$. By definition of $\sup A$, for each $n \in \mathbb{N}$ there exists $x_n \in A$ with $\sup A - \frac{1}{n} < x_n \leq \sup A$. If we construct a sequence $(x_n)_{n\in\mathbb{N}}$ in A in this way, it follows from the squeeze theorem that $\lim_{n\to\infty} x_n = \sup A$. Since A is closed, this shows that $\sup A \in A$. Analogously, one shows that $\inf A \in A$, so by the above reasoning, $\max f = \sup A$ and $\max f = \inf A$ both exist. □

Remark 3.33. (1) The hypothesis that K is closed is necessary in Theorem 3.32, see the example of Remark 3.31.(1).

(2) The hypothesis that K is bounded is necessary in Theorem 3.32: consider $f : [0, +\infty) \to \mathbb{R}$, $f(x) = \frac{x}{x+1}$. Here, $[0, +\infty)$ is closed and f is continuous, but f has no maximum since $\sup f([0, +\infty)) = 1$ which is nowhere attained.

(3) The hypothesis that f is continuous is necessary in Theorem 3.32: Consider

$$f : [-1,1] \to \mathbb{R}, \qquad f(x) = \begin{cases} |x| & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

Here, $[0,1]$ is compact, but f does not have a minimum, since $f([-1,1]) = (0,1]$.

Theorem 3.34. *Let $K \subset \mathbb{C}$ be compact and $f : K \to \mathbb{C}$ be continuous and injective. Then $f^{-1} : f(K) \to K$ is continuous.*

Proof. (This proof is not discussed in the lecture videos.)

This is shown similar to Theorem 3.23. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence in $f(K)$ which converges to $y_0 \in f(K)$. Then $(f^{-1}(y_n))_{n\in\mathbb{N}}$ is a sequence in K. Since K is bounded, $(f^{-1}(y_n))_{n\in\mathbb{N}}$ has a convergent subsequence $(f^{-1}(y_{n_k}))_{k\in\mathbb{N}}$. Since K is closed, $x_0 := \lim_{k\to\infty} f^{-1}(y_{n_k}) \in K$. Since f is continuous, it follows that

$$y_0 = \lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(f^{-1}(y_{n_k})) = f(x_0).$$

Since f is injective, it follows that $f^{-1}(y_0) = x_0$. Since this is true for *every* convergent subsequence, it follows in analogy with the proof of Theorem 3.23, it follows that $\lim_{n\in\mathbb{N}} f^{-1}(y_n) = f^{-1}(y_0)$. It follows that f^{-1} is continuous. □

Definition 3.35. Let $D \subset \mathbb{C}$. A function $f : D \to \mathbb{C}$ is called *uniformly continuous* if

$$\forall \varepsilon > 0\ \exists \delta > 0 : \quad |f(x) - f(y)| < \varepsilon\ \ \forall x, y \in D \text{ with } |x - y| < \delta.$$

The definition of uniform continuity looks very similar to the one of continuity itself, so what is the difference between these two notions? In the definition of continuity via ε's and δ's, we have defined it *in a point*. We have fixed one particular point and demand that for each $\varepsilon > 0$ there exists a suitable $\delta > 0$ *that might depend on that point*. For a uniformly continuous function we demand that given $\varepsilon > 0$, one can choose *the same δ for any point in D*. This shows in particular that every uniformly continuous function is continuous.
To illustrate the difference, we discuss an example of a function that is continuous, but not uniformly continuous.

Example 3.36. The function $f : (0,1] \to \mathbb{R}$, $f(x) = \frac{1}{x}$, is not uniformly continuous. To show this, we need to find an $\varepsilon_0 > 0$, such that for each $\delta > 0$ there are $x, y \in (0,1]$ with $|x - y| < \delta$, but $|f(x) - f(y)| \geq \varepsilon_0$. We want to prove this for $\varepsilon_0 = \frac{1}{2}$. We first compute that

$$|f(x) - f(y)| = \left|\frac{1}{x} - \frac{1}{y}\right| = \frac{|x-y|}{xy} \qquad \forall x, y \in (0,1].$$

Given $\delta > 0$, let $n \in \mathbb{N}$ with $\frac{1}{n} < \delta$. Let $x = \frac{1}{n}$ and $y = \frac{2}{n}$. Then $|x - y| < \delta$, but

$$|f(x) - f(y)| = \frac{\frac{1}{n}}{\frac{1}{n}\cdot\frac{2}{n}} = \frac{n}{2} \geq \frac{1}{2}.$$

Hence, f is not uniformly continuous.

See the corresponding lecture video for a graphical illustration of this example.

Remark 3.37. Every Lipschitz-continuous function is uniformly continuous. This follows as in the proof of Theorem 3.10, since for each $\varepsilon > 0$ one can choose $\delta = \frac{\varepsilon}{L}$, where L is a Lipschitz constant for the function.

Theorem 3.38. *Let $K \subset \mathbb{C}$ be compact and let $f : K \to \mathbb{C}$ be continuous. Then f is uniformly continuous.*

Proof. Assume that f was not uniformly continuous. Then there exists $\varepsilon_0 > 0$ such that for each $\delta > 0$ there exist $x, y \in K$ with $|x - y| < \delta$ and $|f(x) - f(y)| \geq \varepsilon_0$. In particular, for each $n \in \mathbb{N}$ there exist $x_n, y_n \in K$ with $|x_n - y_n| < \delta$ and $|f(x_n) - f(y_n)| \geq \varepsilon_0$. Since K is compact, hence bounded, $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$. Put $a := \lim_{k\to\infty} x_{n_k}$. Since K is closed, $a \in K$. Since $\lim_{k\to\infty} |x_{n_k} - y_{n_k}| \leq \lim_{k\to\infty} \frac{1}{n_k} = 0$, it follows that $\lim_{k\to\infty} y_{n_k} = a$ as well. But since $|f(x_{n_k}) - f(x_{n_k})| \geq \varepsilon_0$ for all $k \in \mathbb{N}$, the sequences $(f(x_{n_k}))_{k\in\mathbb{N}}$ and $(f(y_{n_k}))_{k\in\mathbb{N}}$ can not converge to the same value, such that f is not continuous in a. ↯
This contradicts the assumption on f, hence f is uniformly continuous. □

3.4 Limits of continuous functions

From its definition one can see that continuity is a *local* notion, i.e. the continuity of a function at a point a depends only on its behaviour at points which are close to a and is independent of all other points. We will make this statement formally precise in Theorem 3.42 which requires some preparations.

Definition 3.39. Let $D \subset \mathbb{C}$ and $x_0 \in D$. A subset $A \subset D$ is called *a D-neighborhood of x_0 if there exists an $r > 0$*, such that

$$B_r(x_0) \cap D \subset A.$$

where $B_r(x_0) := \{x \in D \mid |x - x_0| < r\}$.

Remark 3.40. (1) $B_r(x_0)$ is the set of points inside a circle of radius r centered at x_0. It differs from the set $D_r(x_0)$ which we introduced in Theorem 3.28 in that the boundary of the circle, i.e. the circle line[4], does not belong to $B_r(x_0)$, but to $D_r(x_0)$.

(2) The sets $B_r(x_0)$ indirectly appeared in the definition of continuity. Namely, by definition $f : D \to \mathbb{C}$ is by definition continuous in a if

$$\forall \varepsilon > 0 \;\; \exists \delta > 0 : \;\; |f(x) - f(a)| < \varepsilon \quad \forall x \in B_\delta(a) \cap D.$$

Definition 3.41. Let X and Y be sets, let $f : X \to Y$ be a map and let $A \subset X$.

a) The map

$$f|_A : A \to Y, \qquad f|_A(x) = f(x) \quad \forall x \in A,$$

is called *the restriction of f to A.*

b) Let $g : A \to Y$ be a map. f is called *an extension of g to X* if $f(x) = g(x)$ for all $x \in A$, i.e. if $f|_A = g$.

The local character of continuity is expressed in the following theorem. Roughly stated, it says that continuity of a map in a point is equivalent to the continuity of the restriction to a subset at the point under the assumption that the subset contains all point of the domain that are "very close" to that point.

[4]Not to be confused with the London underground line of the same name. (Sorry.)

Theorem 3.42. *Let $D \subset \mathbb{C}$ and $f : D \to \mathbb{C}$. Let $A \subset D$ and $a \in A$.*

a) If f is continuous in a, then $f|_A : A \to \mathbb{C}$ is continuous in a.

b) Assume that A is a D-neighborhood of a. Then f is continuous in a if and only if $f|_A$ is continuous in a.

Proof. a) Let $\varepsilon > 0$. Since f is continuous, there exists $\delta > 0$, such that $|f(x) - f(a)| < \varepsilon$ for all $x \in B_\delta(a) \cap D$. Since $A \subset D$, this implies that

$$|f|_A(x) - f|_A(a)| = |f(x) - f(a)| < \varepsilon \quad \forall x \in B_\delta(a) \cap A.$$

Since ε was arbitrary, this shows the continuity of $f|_A$ in a.

b) "$\Rightarrow$" follows from part a), we only need to show "$\Leftarrow$".

Since A is a D-neighborhood of a, there exists $r > 0$, such that $B_r(a) \cap D \subset A$. In particular, $B_r(a) \cap A = B_r(a) \cap D$ and thus $B_{r'}(a) \cap A = B_{r'}(a) \cap D$ for all $r' \in (0, r]$.

Let $\varepsilon > 0$. Since $f|_A$ is continuous in a, there exists $\delta > 0$ with

$$|f(x) - f(a)| = |f|_A(x) - f|_A(a)| < \varepsilon \qquad \forall x \in B_\delta(a) \cap A.$$

Now let $\delta' := \min\{\delta, r\} > 0$. Since $\delta' \leq \delta$, it particularly follows that $|f(x) - f(a)| < \varepsilon$ for all $x \in B_{\delta'}(a) \cap A$. Since $\delta' \leq r$, it follows that $B_{\delta'}(a) \cap A = B_{\delta'}(a) \cap D$, such that

$$|f(x) - f(a)| < \varepsilon \qquad \forall x \in B_{\delta'}(a) \cap D.$$

Since ε was chosen arbitrarily, this shows that f is continuous in a.

□

Corollary 3.43. *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series of radius of convergence ρ with $\rho \neq 0$. Then $f : B_\rho(0) \to \mathbb{C}$, $f(x) = \sum_{n=0}^{\infty} a_n x^n$, is continuous. Here, $B_{+\infty}(0) := \mathbb{C}$.*

Proof. Let $a \in B_\rho(0)$, such that $|a| < \rho$ and let $R \in (|a|, \rho)$. With $r := R - |a| > 0$, it holds for all $z \in B_r(a)$ by the triangle inequality that

$$|z| \leq |z - a| + |a| < r + |a| = R,$$

which shows that $B_r(a) \subset D_R$. Hence, D_R is a $\mathbb{C}$-neighborhood of a. By Theorem 3.13, $f|_{D_R}$ is Lipschitz-continuous, hence continuous, so it follows from Theorem 3.42.b) that f is continuous in a. □

Next we want to tackle extensions of continuous functions. Given $f : D \to \mathbb{C}$ and $a \in \mathbb{C} \setminus D$, we can of course define an extension

$$\tilde{f} : D \cup \{a\} \to \mathbb{C}$$

of f as a map by simply assigning an arbitrary value of $\mathbb{C}$. This gives rise to the following question:

Can we find an extension of f to $D \cup \{a\}$ *which is continuous in a?*

If a does not "touch" D in a sense that we will make precise momentarily, it is easy to see that *any* extension of f to $D \cup \{a\}$ is continuous, but if a "touches" D, such a continuous extension must not necessarily exist. (We shall see some counterexamples below.)
Before studying this question, we need formalize the notion of points "touching" subsets of $\mathbb{C}$.

Definition 3.44. Let $A \subset \mathbb{C}$.

(1) Let $a \in \mathbb{C}$. We call a an *adherent point of* A if for each $\varepsilon > 0$ there exists an $x \in A$ with $|x - a| < \varepsilon$, i.e. if $B_\varepsilon(a) \cap A \neq \varnothing$ for each $\varepsilon > 0$

(2) The set
$$\overline{A} := \{a \in \mathbb{C} \mid a \text{ is an adherent point of } A\}$$
is called *the closure of* A.

Remark 3.45. (1) Every point in A is an adherent point of A, i.e. $A \subset \overline{A}$, since $B_\varepsilon(a) \cap A \supset \{a\} \neq \varnothing$ for all $a \in A$ and $\varepsilon > 0$. Conversely, not every adherent point a of A must lie in A, see e.g. $A = (0, 1]$ and $a = 0$.

(2) a is an adherent point in A if and only if there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in A which converges to a. From this fact, one derives that $\overline{A} = A$ if and only if A is a closed subset of $\mathbb{C}$.

(3) Let $f : D \to \mathbb{C}$ be continuous and let $a \in \mathbb{C} \setminus D$. If a is *not* an adherent point of D, then *every* extension $g : D \cup \{a\} \to \mathbb{C}$ of f is continuous.

If a is an adherent point, then in general there might be no continuous extension of f to $D \cup \{a\}$ at all. A simple example is given by
$$f : \mathbb{R} \setminus \{0\} \to \mathbb{R}, \qquad f(x) = \frac{1}{x}.$$
Here, 0 is an adherent point of $\mathbb{R} \setminus \{0\}$, but there can be no continuous extension of f since e.g. $\lim_{n\to\infty} f(\frac{1}{n}) = +\infty$.

Theorem 3.46. *Let $D \subset \mathbb{C}$, let $f : D \to \mathbb{C}$ be continuous and let $a \in \overline{D}$. If there exists a continuous extension $g : D \cup \{a\} \to \mathbb{C}$ of f, then it will be unique.*

Proof. If $a \in D$, then $g = f$, so the statement is obvious. Assume that $a \notin D$. Let $g_1, g_2 : D \cup \{a\} \to \mathbb{C}$ both be continuous extensions of f. Then $h : D \cup \{a\} \to \mathbb{C}$, $h(x) := g_1(x) - g_2(x)$ is continuous. Moreover,
$$h(x) = \begin{cases} 0 & \text{if } x \in D, \\ c & \text{if } x = a, \end{cases}$$
where $c := g_1(a) - g_2(a)$. Since h is continuous in a,
$$\forall \varepsilon > 0 \ \ \exists \delta > 0 : \ \ |h(x) - h(a)| < \varepsilon \quad \forall x \in B_\delta(a) \cap D.$$
Here, $B_\delta(a) \cap D \neq \varnothing$ since a is an adherent point of D. Since $h(x) = 0$ for all $x \in D$, this implies that $|0 - c| < \varepsilon$, i.e. $|c| < \varepsilon$ for each $\varepsilon > 0$. This shows that $|c| = 0$ and thus $c = 0$, i.e. $g_1(a) = g_2(a)$. □

Definition 3.47. Let $D \subset \mathbb{C}$, $a \in \overline{D}$ and $f : D \to \mathbb{C}$ be continuous. We say that f *has a limit in* a if there exists a continuous extension $g : D \cup \{a\} \to \mathbb{C}$ of f. In this case, we write
$$\lim_{x\to a} f(x) := g(a).$$
(By Theorem 3.46, this expression is well-defined.)

Example 3.48. (1) Let $f : [0,1) \cup (1,+\infty) \to \mathbb{R}$, $f(x) = \frac{\sqrt{x}-1}{x-1}$. We compute for all $x \in \mathbb{R} \setminus \{1\}$ that

$$\frac{\sqrt{x}-1}{x-1} = \frac{\sqrt{x}-1}{(\sqrt{x}+1)(\sqrt{x}-1)} = \frac{1}{\sqrt{x}+1}.$$

Hence, $g : [0,+\infty) \to \mathbb{R}$, $g(x) = \frac{1}{\sqrt{x}+1}$, which is continuous, satisfies $g(x) = f(x)$ for all $x \in \mathbb{R} \setminus \{1\}$ and we obtain

$$\lim_{x\to 1} \frac{\sqrt{x}-1}{x-1} = g(1) = \frac{1}{\sqrt{1}+1} = \frac{1}{2}.$$

(2) Let $n \in \mathbb{N}$ and let $f_n : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $f_n(x) = \frac{(1+x)^n-1}{x}$. We want to show that each f_n has a limit in 0 and compute it. For every $x \in \mathbb{R}$, the binomial theorem implies

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

Using this identity, we compute that

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

where we performed an index shift in the final equality. Thus, the polynomial $g_n : \mathbb{R} \to \mathbb{R}$, $g_n(x) = \sum_{k=0}^{n-1} \binom{n}{k+1} x^k$, is a continuous extension of f_n, which shows that

$$\lim_{x\to 0} f_n(x) = g_n(0) = \binom{n}{1} = \frac{n!}{(n-1)!} = n \qquad \forall n \in \mathbb{N}.$$

The following theorem provides a justification for calling such a number the *limit* of a function.

Theorem 3.49. *Let $D \subset \mathbb{C}$, $a \in \overline{D}$, $f : D \to \mathbb{C}$ be continuous and $L \in \mathbb{C}$. The following are equivalent:*

(i) $\lim_{x\to a} f(x) = L$.

(ii) For each sequence $(x_n)_{n\in\mathbb{N}}$ in D with $\lim_{n\to\infty} x_n = a$, it holds that $\lim_{n\to\infty} f(x_n) = L$.

Proof. This follows immediately from applying Theorem 3.3 to a continuous extension of f to $D \cup \{a\}$. □

The previous theorem enables us to transfer computational rules for sequences and continuous functions to the limits introduced in thei section.

Theorem 3.50. *Let $D \subset \mathbb{C}$ and $a \in \overline{D}$. Let $f, g : D \to \mathbb{C}$ be functions which both have limits in a. Then:*

a) $f + g : D \to \mathbb{C}$ has a limit in a with

$$\lim_{x\to a}(f+g)(x) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x).$$

b) $f \cdot g : D \to \mathbb{C}$ has a limit in a with

$$\lim_{x\to a}(f\cdot g)(x) = \Big(\lim_{x\to a} f(x)\Big) \cdot \Big(\lim_{x\to a} g(x)\Big).$$

c) If $\lim_{x\to a} g(x) \neq 0$*, then there exists a* $\overline{D}$*-neighborhood* A *of* a *with* $g(x) \neq 0$ *for all* $x \in A \cap D$ *and* $\frac{f}{g} : A \cap D \to \mathbb{C}$, $\frac{f}{g}(x) := \frac{f(x)}{g(x)}$ *has a limit in* a *with*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}.$$

Proof. This follows by applying Theorems 3.5 and 3.49 to the continuous extensions of f and g to $D \cup \{a\}$. □

In the following theorem, we discuss a criterion for the existence of the limit of a function in an adherent point without having to name an explicit number as the limit.

Theorem 3.51. *Let* $D \subset \mathbb{C}$*, let* $a \in \overline{D}$ *and let* $f : D \to \mathbb{C}$ *be continuous.* f *has a limit in* a *if and only if*

$$\forall \varepsilon > 0\ \exists \delta > 0 : \quad |f(x) - f(y)| < \varepsilon \ \ \forall x, y \in B_\delta(a) \cap D.$$

Proof. "⇒": Let $g : D \cup \{a\}$ be a continuous extension of f and let $\varepsilon > 0$. Since g is continuous, there exists $\delta > 0$ such that $|g(x) - g(a)| < \frac{\varepsilon}{2}$ for all $x \in B_\delta(a) \cap (D \cup \{a\})$. In particular, for all $x \in B_\delta(a) \cap D$, this implies that $|f(x) - G(a)| < \frac{\varepsilon}{2}$. Thus, by the triangle inequality,

$$|f(x) - f(y)| = |f(x) - g(a) - (f(y) - g(a))| \leq |f(x) - g(a)| + |f(y) - g(a)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $x, y \in B_\delta(a) \cap D$. This is what we wanted to show.
"⇐": Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in D with $\lim_{n\to\infty} x_n = a$.
Let $\varepsilon > 0$ and let $\delta > 0$, such that $|f(x) - f(y)| < \varepsilon$ for all $x, y \in B_\delta(a) \cap D$. Since $(x_n)_{n\in\mathbb{N}}$ converges to a, there exists $n_0 \in \mathbb{N}$ such that $x_n \in B_\delta(a) \cap D$s for all $n \geq n_0$. By our choice of delta, this implies

$$|f(x_n) - f(x_m)| < \varepsilon \qquad \forall m, n \geq n_0.$$

Since ε was chosen arbitrarily, we have shown that $(f(x_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}$, which converges by Corollary 2.52. Put $b := \lim_{n\to\infty} f(x_n)$. We claim that the value of b is independent of the choice of $(x_n)_{n\in\mathbb{N}}$. Namely, if $(x'_n)_{n\in\mathbb{N}}$ is another sequence in D converging to a define a third sequence $(y_n)_{n\in\mathbb{N}}$ as $(x_1, x'_1, x_2, x'_2, \dots)$, i.e. by

$$y_n := \begin{cases} x_k & \text{if } n = 2k-1, \\ x'_k & \text{if } n = 2k. \end{cases}$$

Then $(y_n)_{n\in\mathbb{N}}$ converges to a as well and the same argument as for $(x_n)_{n\in\mathbb{N}}$ shows that $(f(y_n))_{n\in\mathbb{N}}$ converges. Since both $(x_n)_{n\in\mathbb{N}}$ and $(x'_n)_{n\in\mathbb{N}}$ are subsequences of $(y_n)_{n\in\mathbb{N}}$, it follows that

$$\lim_{n\to\infty} x'_n = \lim_{n\to\infty} y_n = \lim_{n\to\infty} x_n = b.$$

Thus, it follows from Theorem 3.49 that $\lim_{x\to a} f(x) = b$. □

Corollary 3.52. *Let* $D \subset \mathbb{C}$*. If* $f : D \to \mathbb{C}$ *is uniformly continuous, then there exists a continuous extension* $g : \overline{D} \to \mathbb{C}$ *of* f *to the closure* $\overline{D}$ *of* D*.*

Proof. Let $a \in \mathbb{C}$ be an adherent point of D and let $\varepsilon > 0$. Since f is uniformly continuous, there exists $\delta > 0$, such that $|f(x) - f(y)| < \varepsilon$ for all $x, y \in D$ with $|x - y| < 2\delta$. Since for all $x, y \in B_\delta(a) \cap D$ it holds that $|x - y| \leq |x - a| + |y - a| < \delta + \delta = 2\delta$, this shows that

$$|f(x) - f(y)| < \varepsilon \quad \forall x, y \in B_\delta(a) \cap D.$$

By Theorem 3.51, this shows that f has a limit in a. Since f has a limit in every adherent point of D, this shows that f has a continuous extension to $\overline{D}$. □

In the following we want to study limits of functions whose domains are subsets of $\mathbb{R}$. Here, we can study several phenomena that specifically occur in the real setting. For the first such phenomenon, we take a look at a motivating example. Let

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} x & \text{if } x < 0, \\ x+1 & \text{if } x \geq 0. \end{cases}$$

Obviously, f is not continuous in 0. However, the restriction $f|_{[0,+\infty)}$ is continuous (while $f|_{(-\infty,0]}$ is not). In other words, we obtain continuity in 0 if we only approach 0 "from the right-hand side". This is a general phenomenon happening for real functions which are continuous up to certain "jumps" and we shall formalize these notions in the following definition.

Definition 3.53. Let $D \subset \mathbb{R}$, let $f : D \to \mathbb{C}$ be a map and let $a \in \mathbb{R}$. Put $D_- := D \cap (-\infty, a)$ and $D_+ := D \cap (a, +\infty)$.

a) *f has a left-sided limit in a* if a is an adherent point of D_- and if $f|_{D_-}$ has a limit in a. We then write
$$\lim_{x \nearrow a} f(x) := \lim_{x \to a} f|_{D_-}(x).$$

b) *f has a right-sided limit in a* if a is an adherent point of D_+ and if $f|_{D_+}$ has a limit in a. We then write
$$\lim_{x \searrow a} f(x) := \lim_{x \to a} f|_{D_+}(x).$$

c) Assume that $a \in D$. We call *f left-continuous in a* if it has a left-sided limit in a with $\lim\limits_{x \nearrow a} f(x) = f(a)$.

d) Assume that $a \in D$. We call *f right-continuous in a* if it has a right-sided limit in a with $\lim\limits_{x \searrow a} f(x) = f(a)$.

Example 3.54. The Heaviside function from Example 3.4 is right-continuous in 0, but not left-continuous in 0.

Theorem 3.55. *Let $D \subset \mathbb{R}$, $a \in D$ and $f : D \to \mathbb{C}$. The following are equivalent:*

(i) f is continuous in a.

(ii) f is left-continuous in a and right-continuous in a.

Proof. (This proof is not discussed in the lecture videos.)

(i) $\Rightarrow$ (ii): This is obvious.
(ii) $\Rightarrow$ (i): Assume that f was not continuous in a. Then there exists $(x_n)_{n\in\mathbb{N}}$ be a sequence in D with $\lim_{n\to\infty} x_n = a$, but such that $(f(x_n))_{n\in\mathbb{N}}$ does not converge to $f(a)$. Then there exists $\varepsilon_0 > 0$, such that the set $M := \{n \in \mathbb{N} \mid |f(x_n) - f(a)| \geq \varepsilon_0\}$ is infinite. Then at least one of the sets $M \cap (-\infty, a]$ and $M \cap [a, +\infty)$ is infinite. If $M \cap [a, +\infty)$ is infinite, then there exists a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ with $x_{n_k} \geq a$ for all $k \in \mathbb{N}$, $\lim_{k\in\mathbb{N}} x_{n_k} = a$, but $|f(x_{n_k}) - f(a)| \geq \varepsilon_0$ for all $k \in \mathbb{N}$. This is a contradiction to right-continuity, hence f must be continuous in a. If $M \cap (-\infty, a]$ is infinite, the very same line of argument shows that left-continuity is violated, hence f is continuous in a as well. □

We want to study the situation that a function $f(x)$ in a real variable approaches or tends to some value as x tends to $+\infty$ or to $-\infty$. We will extend the terminology of limits of functions to this case.

Definition 3.56. Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{C}$.

a) Assume that D is not bounded from above. We say that f *has a limit in* $+\infty$ if there exists an $L \in \mathbb{C}$, such that

$$\forall \varepsilon > 0 \; \exists R \in \mathbb{R} : \;\; |f(x) - L| < \varepsilon \;\; \forall x \in D \cap (R, +\infty).$$

In this case we write $\lim\limits_{x \to +\infty} f(x) := L$.

b) Assume that D is not bounded from below. We say that f *has a limit in* $-\infty$ if there exists an $L \in \mathbb{C}$, such that

$$\forall \varepsilon > 0 \; \exists R \in \mathbb{R} : \;\; |f(x) - L| < \varepsilon \;\; \forall x \in D \cap (-\infty, R).$$

In this case we write $\lim\limits_{x \to -\infty} f(x) := L$.

Example 3.57. (1) Let $q \in \mathbb{Q}$ with $q > 0$ and let $f : (0, +\infty) \to \mathbb{R}$, $f(x) = \frac{1}{x^q}$.

Let $\varepsilon > 0$. By Corollary 1.53, there exists $N \in \mathbb{N}$ with $\frac{1}{N} < \varepsilon^{1/q}$ and thus $\frac{1}{x} \leq \frac{1}{N} < \varepsilon^{1/q}$ for each $x \in [N, +\infty)$. This yields that

$$|f(x)| = \frac{1}{x^q} < \varepsilon \qquad \forall x \in (N, +\infty).$$

Since ε was arbitrary, this shows that $\lim\limits_{x \to +\infty} \dfrac{1}{x^q} = 0$.

(2) Let $n \in \mathbb{N}$ and consider[5] $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $f(x) = \frac{1}{x^n}$.

For $\varepsilon > 0$ let $N \in \mathbb{N}$ be chosen as in (1) with $\frac{1}{N} < \varepsilon^{1/n}$. For each $x < -N$ it holds that $-x > N$ and thus $-\frac{1}{x} < \frac{1}{N}$. For n odd, this yields that $-\frac{1}{x^n} < \frac{1}{N^n}$, while for n even, $\frac{1}{x^n} < \frac{1}{N^n}$. In both cases, we derive

$$|f(x)| = \left| \frac{1}{x^n} \right| =< \frac{1}{N^n} < \varepsilon \qquad \forall x \in (-\infty, -N).$$

This shows that $\lim\limits_{x \to -\infty} \dfrac{1}{x^n} = 0$.

Remark 3.58. (1) The computational rules from Theorem 3.50 are valid for limits in $+\infty$ or $-\infty$ as well. (We will use this observation without proving it explicitly.)

(2) The previous definition generalizes the notion of the limit of a sequence. By definition, a complex sequence is a map $a : \mathbb{N} \to \mathbb{C}$ and if we construct the limit from part a) of Definition 3.56 for $D = \mathbb{N}$, one easily realizes that $\lim_{x \to +\infty} a(x) = \lim_{n \to \infty} a_n$, i.e. that the limit of a in $+\infty$ coincides with the limit of $(a_n)_{n \in \mathbb{N}}$ (in case they exist).

[5]Note that $\frac{1}{x^q}$ is not well-defined for negative x if $q \notin \mathbb{N}$ since for $q = \frac{m}{n}$, $x^q = (\sqrt[n]{x})^m$ and n-th roots are only well-defined for non-negative numbers.

Theorem 3.59. *Let $D \subset \mathbb{R}$, $f : D \to \mathbb{C}$ and $L \in \mathbb{C}$.*

a) Assume that D is not bounded from above. Then $\lim\limits_{x\to+\infty} f(x) = L$ if and only if for every sequence $(x_n)_{n\in\mathbb{N}}$ in D with $\lim\limits_{n\to\infty} x_n = +\infty$ it holds that $\lim\limits_{n\to\infty} f(x_n) = L$.

b) Assume that D is not bounded from below. Then $\lim\limits_{x\to-\infty} f(x) = L$ if and only if for every sequence $(x_n)_{n\in\mathbb{N}}$ in D with $\lim\limits_{n\to\infty} x_n = -\infty$ it holds that $\lim\limits_{n\to\infty} f(x_n) = L$.

Proof. (This proof is not discussed in the lecture videos.)

We will only proof part a), since part b) is shown completely analogously, replacing right-sided limits by left-sided limits in the line of argument below.

"$\Rightarrow$": This follows immediately from the definition of $\lim\limits_{x\to\infty} f(x)$.

"$\Leftarrow$": **Claim:** Let $D' := \{x \in \mathbb{C}^* \mid \frac{1}{x} \in D\}$. Then f has a limit in $+\infty$ if and only if

$$g : D' \to \mathbb{C}, \qquad g(x) = f(\tfrac{1}{x}),$$

has a right-sided limit in 0. In case of existence, the two limits coincide.

Proof of claim. "$\Rightarrow$": Put $L := \lim_{x\to+\infty} f(x)$ and let $\varepsilon > 0$. By assumption, there eixsts $R > 0$ with $|f(x) - L| < \varepsilon$ for all $x \in (R, +\infty) \cap D$. Putting $\delta := \frac{1}{R} > 0$, this implies that

$$|g(x) - L| = \Big|f\Big(\frac{1}{x}\Big) - L\Big| < \varepsilon \qquad \forall x \in (0, \delta) \cap D'.$$

Since $\varepsilon > 0$ was arbitrary, this shows that $\lim_{x\searrow 0} g(x) = L$.
"$\Leftarrow$": Put $L := \lim_{x\searrow 0} g(x)$. Then for all $\varepsilon > 0$ there exists $\delta > 0$ with $|g(x) - L| < \varepsilon$ for all $x \in (0, \delta) \cap D'$. Putting $R := \frac{1}{\delta}$, this implies that

$$|f(x) - L| = \Big|g\Big(\frac{1}{x}\Big) - L\Big| < \varepsilon \qquad \forall x \in (R, +\infty) \cap D,$$

which shows that $\lim_{x\to+\infty} f(x) = L$ ■

Let $(x_n)_{n\to\infty}$ be a sequence in $D \cap (0, +\infty)$ which converges to 0. Then by Theorem 2.43, $(\frac{1}{x_n})_{n\in\mathbb{N}}$ diverges to $+\infty$ and by assumption

$$\lim_{n\to\infty} g(x_n) = \lim_{n\to\infty} f\Big(\frac{1}{x_n}\Big) = L.$$

Since this holds for *any* such sequence $(x_n)_{n\in\mathbb{N}}$, it follows from Theorem 3.49 that $\lim\limits_{x\searrow 0} g(x) = L$ and by the claim proven above, this yields

$$\lim_{x\to+\infty} f(x) = \lim_{x\searrow 0} g(x) = L.$$

□

We want to take a look at real functions. Similar as for real sequences, if inequalities between functions hold in a neighborhood of the point at which the limits, then this will lead to inequalities between the limits.

Theorem 3.60. *Let $D \subset \mathbb{R}$ and $f, g : D \to \mathbb{R}$.*

a) Let $a \in \overline{D}$ and assume that both f and g have limits in a. If there exists a D-neighborhood A of a such that $f(x) \leq g(x)$ for all $x \in A$, then

$$\lim_{x \to a} f(x) \leq \lim_{x \to a} g(x).$$

b) Assume that D is not bounded from above and that both f and g have limits in $+\infty$. If there exists an $R \in \mathbb{R}$, such that $f(x) \leq g(x)$ for all $x \in D \cap (R, +\infty)$, then

$$\lim_{x \to +\infty} f(x) \leq \lim_{x \to +\infty} g(x).$$

c) Assume that D is not bounded from below and that both f and g have limits in $-\infty$. If there exists an $R \in \mathbb{R}$, such that $f(x) \leq g(x)$ for all $x \in D \cap (-\infty, R)$, then

$$\lim_{x \to -\infty} f(x) \leq \lim_{x \to -\infty} g(x).$$

Proof. This follows from combining Theorem 2.26 about inequalities between sequences and their limits with Theorems 3.49 and 3.59 describing limits of functions in terms of sequences. We omit the details. □

Remark 3.61. In the same way that we derived the squeeze theorem for sequences (Theorem 2.28) from Theorem 2.26, one can derive a squeeze theorem for function from Theorem 3.60. We refrain from giving the precise statement, but leave its formulation as a little exercise to the reader.

We conclude this section by discussing so-called improper limits of functions. They are formalizing the notion of a function tending to $+\infty$ or to $-\infty$ when approaching an adherent point which is not in the domain of f. While parts a) and b) of the following definition could be formulated for functions with domains in $\mathbb{C}$ as well, we shall stick with the real case for simplicity.

Definition 3.62. Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$.

a) Let $a \in \overline{D} \setminus D$. We say that *$f$ has the improper limit $+\infty$ in a* if

$$\forall R \in \mathbb{R}\ \exists \delta > 0 :\ \ f(x) > R\ \ \forall x \in B_\delta(a) \cap D.$$

In this case, we write $\lim_{x \to a} f(x) = +\infty$.

b) Let $a \in \overline{D} \setminus D$. We say that *$f$ has the improper limit $-\infty$ in a* if

$$\forall R \in \mathbb{R}\ \exists \delta > 0 :\ \ f(x) < R\ \ \forall x \in B_\delta(a) \cap D.$$

In this case, we write $\lim_{x \to a} f(x) = -\infty$.

c) Assume that D is not bounded from above. We say that *f has the improper limit $+\infty$ in $+\infty$* if

$$\forall R \in \mathbb{R}\ \exists C \in \mathbb{R} :\ \ f(x) > R\ \ \forall x \in (C, +\infty) \cap D.$$

In this case, we write $\lim_{x \to +\infty} f(x) = +\infty$.

d) Assume that D is not bounded from below. We say that *f has the improper limit $-\infty$ in $+\infty$* if

$$\forall R \in \mathbb{R}\ \exists C \in \mathbb{R}: \quad f(x) < R \ \ \forall x \in (C, +\infty) \cap D.$$

In this case, we write $\lim\limits_{x\to+\infty} f(x) = -\infty$.

In analogy with c) and d), one defines improper limits in $-\infty$. We omit the details.

Example 3.63. (1) Let $q \in \mathbb{Q}$ with $q > 0$ and let $f : (0, +\infty) \to \mathbb{R}$, $f(x) = \frac{1}{x^q}$. We want to investigate $\lim_{x\to 0} f(x)$.

Let $C \in \mathbb{R}$. Then by Corollary 1.53 there exists an $N \in \mathbb{N}$ with $\frac{1}{N} < \frac{1}{C^{1/q}}$. Put $\delta := \frac{1}{N}$. Then $x < \frac{1}{C^{1/q}}$ for all $x \in (0, \delta)$. We compute that

$$x < \frac{1}{C^{1/q}} \quad \Leftrightarrow \quad x^q < \frac{1}{C} \quad \Leftrightarrow \quad \frac{1}{x^q} > C.$$

Hence, we have shown that

$$\frac{1}{x^q} > C \qquad \forall x \in (0, \delta) = B_\delta(0) \cap (0, +\infty),$$

which shows that $\lim_{x\to 0} f(x) = +\infty$.

(2) For every $k \in \mathbb{N}$, it holds that $\lim\limits_{x\to-\infty}(x^{2k}) = \lim\limits_{x\to+\infty}(x^{2k}) = +\infty$, while $\lim\limits_{x\to-\infty} x^{2k-1} = -\infty$ and $\lim\limits_{x\to+\infty} x^{2k-1} = +\infty$ for each $k \in \mathbb{N}$. We leave the proof of these facts as an exercise to the reader.

3.5 The exponential function, logarithms and powers

In this section and the following one, we want to revisit the exponential series introduced in Section 2.5 as a power series that converges on all of $\mathbb{C}$. In this section, we want to view focus on its restriction to $\mathbb{R}$. Using properties of this restriction, we will introduce a way of defining powers with arbitrary real or complex exponents, generalizing our definition of powers with rational exponents as we shall see below.

Definition 3.64. The function $\exp : \mathbb{C} \to \mathbb{C}$, $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, is called *the exponential function*.

By Corollary 3.43, exp is continuous. We recall that we have shown in Theorem 2.99 that

$$\exp(x + y) = \exp(x) \cdot \exp(y) \qquad \forall x, y \in \mathbb{C}. \tag{3.2}$$

For each $x \in \mathbb{R}$ we put $e^x := \exp(x)$. By Theorem 2.99.e), this does not contradict the previous notation. We want to investigate the behaviour of e^x as x tends to $+\infty$ or to $-\infty$.

Theorem 3.65. *For each $n \in \mathbb{N}_0$, it holds that*

$$\lim_{x\to+\infty} \frac{e^x}{x^n} = +\infty, \qquad \lim_{x\to-\infty} x^n e^x = 0.$$

Proof. By definition, for $x > 0$ it holds that $e^x > \frac{x^{n+1}}{(n+1)!}$, so

$$\frac{e^x}{x^n} > \frac{x}{(n+1)!}.$$

Since $\lim_{x\to+\infty} \frac{x}{(n+1)!} = +\infty$, one easily derives that $\lim_{x\to+\infty} \frac{e^x}{x^n} = +\infty$ as well. From the definition of limits of functions, it easily follows that

$$\lim_{x\to-\infty} x^n e^x = \lim_{y\to+\infty} (-y)^n e^{-y} = (-1)^n \lim_{y\to+\infty} \frac{y^n}{e^y}.$$

In analogy with Theorem 2.43, one derives from $\lim_{y\to+\infty} \frac{e^y}{y^n} = +\infty$ that $\lim_{y\to+\infty} \frac{y^n}{e^y} = 0$, which shows the second claim. □

Remark 3.66. Loosely speaking, given any $n \in \mathbb{N}_0$ the first limit in Theorem 3.65 shows that e^x *grows faster to* $+\infty$ *than* x^n *for* $x \to +\infty$.

Theorem 3.67. *a) If* $x \in \mathbb{R}$, *then* $\exp(x) \in (0, +\infty)$.

b) $\exp_{\mathbb{R}} : \mathbb{R} \to (0, +\infty)$, $\exp_{\mathbb{R}}(x) := \exp(x)$, *is continuous, strictly monotonically increasing and bijective.*

Proof. a) If $x \in [0, +\infty)$, then

$$\exp(x) = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!}.$$

Since $\frac{x^n}{n!} \geq 0$ for each $n \in \mathbb{N}$, it follows that $\exp(x) \in [1 + \infty)$. If $x \in (-\infty, 0)$, then by Theorem 2.99.c),

$$\exp(x) = \frac{1}{\exp(-x)} > 0.$$

b) $\exp_{\mathbb{R}}$ is well-defined by a) and continuous since exp itself is continuous. Let $x, y \in \mathbb{R}$ with $x < y$. By (3.2),

$$\exp(y) = \exp(y - x + x) = \exp(y - x) \cdot \exp(x).$$

Since $y - x > 0$, it follows that $\exp(y - x) = 1 + \sum_{n=1}^{\infty} \frac{(y-x)^n}{n!} > 1$, since each of the elements of the series is positive. Hence, $\exp(y) > \exp(x)$, so we have shown that $\exp_{\mathbb{R}}$ is strictly monotonically increasing, which also implies that $\exp_{\mathbb{R}}$ is injective. It only remains to show the surjectivity of $\exp_{\mathbb{R}}$.

By Theorem 3.65, it holds that $\lim_{x\to-\infty} \exp_{\mathbb{R}}(x) = 0$ and $\lim_{x\to+\infty} \exp_{\mathbb{R}}(x) = +\infty$. Hence, for each $y \in (0, +\infty)$ there exist $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$, $\exp(x_1) \leq y$ and $\exp(x_2) \geq y$. By the intermediate value theorem, there exists $\xi \in [x_1, x_2]$ with $\exp(\xi) = y$, hence $\exp_{\mathbb{R}}$ is surjective.

□

Definition 3.68. The *natural logarithm* is given as the inverse function

$$\log := \exp_{\mathbb{R}}^{-1} : (0, +\infty) \to \mathbb{R}.$$

$\log(x)$ is also often denoted by $\ln(x)$.

By Theorem 3.23, $\log : (0,+\infty) \to \mathbb{R}$ is continuous. The basic properties of the exponential function imply properties of the natural logarithm which are summarized in the following theorem.

Theorem 3.69. *a)* $\log 1 = 0$ *and* $\log e = 1$.

b) $\log(x \cdot y) = \log x + \log y$ *for all* $x, y \in (0,+\infty)$.

c) $\log(x^{-1}) = -\log x$ *for all* $x \in (0,+\infty)$.

d) Let $a \in (0,+\infty)$ *and* $q \in \mathbb{Q}$. *Then*

$$\exp(q \cdot \log a) = a^q.$$

Proof. a) This follows from $\exp(0) = 1$ and $\exp(1) = e$.

b) Let $x, y \in (0,+\infty)$. We derive from (3.2) that

$$\exp(\log x + \log y) = \exp(\log x) \cdot \exp(\log y) = x \cdot y.$$

Since $\exp_{\mathbb{R}}$ is bijective and since $\exp(\log(x \cdot y)) = x \cdot y$, this shows the claim.

c) Let $x \in (0,+\infty)$. Using Theorem 2.99.c) we compute that

$$\exp(-\log x) = \exp(\log x)^{-1} = x^{-1}.$$

Arguing as in b), this shows that $-\log x = \log(x^{-1})$.

d) This is shown in complete analogy with Theorem 2.99.e) which discusses the case $a = e$. We omit the details.

□

The observation of Theorem 3.69.d) motivates the following definition of powers whose exponent is an arbitrary real or complex number.

Definition 3.70. Given $a \in (0,+\infty)$ and $x \in \mathbb{C}$, we define $a^x \in \mathbb{C}$ by

$$a^x := \exp(x \cdot \log a).$$

Theorem 3.71. *Let* $a, b \in (0,+\infty)$.

a) $a^0 = 1$.

b) $a^{x+y} = a^x \cdot a^y$ *for all* $x, y \in \mathbb{C}$.

c) $(\frac{1}{a})^x = a^{-x}$ *for all* $x \in \mathbb{C}$.

d) $(ab)^x = a^x \cdot b^x$ *for all* $x \in \mathbb{C}$.

e) $\log(a^x) = x \log a$ *for all* $x \in \mathbb{R}$.

f) $a^{x \cdot y} = (a^x)^y$ *for all* $x, y \in \mathbb{R}$.

Proof. a) By definition, $a^0 = \exp(0 \cdot \log a) = \exp(0) = 1$.

b) We compute that

$$a^{x+y} = \exp((x+y)\log a) = \exp(x \log a + y \log a) \overset{(3.2)}{=} \exp(x \log a) \cdot \exp(y \log a) = a^x \cdot a^y.$$

c) By Theorem 3.69.c), $(\frac{1}{a})^x = \exp(x\log(a^{-1})) = \exp(-x\log(a)) = a^{-x}$.

d) By Theorem 3.69.b),

$$(ab)^x = \exp(x\log(ab)) = \exp(x(\log a + \log b)) = \exp(x\log a + x\log b)$$
$$\overset{(3.2)}{=} \exp(x\log a)\cdot\exp(x\log b) = a^x\cdot b^x.$$

e) By definition, $\exp(x\log a) = a^x = \exp(\log(a^x))$. Since $\exp_{\mathbb{R}}$ is bijective, this implies $\log(a^x) = x\log a$.

f) By e), $(a^x)^y = \exp(y\log(a^x)) = \exp(yx\log a) = a^{xy}$.

□

More generally, logarithms can be defined as inverse maps to maps of the form a^x, where $a > 0$. In the following Theorem/Definition, we summarize the construction and omit its proof which is carried out in analogy with the natural logarithm.

Theorem/Definition 3.72. *For each $a > 0$, the function*

$$f_a : \mathbb{R} \to (0,+\infty), \qquad f(x) = a^x = \exp(x\log a),$$

is continuous, stricly monotonically increasing and bijective. Its inverse function

$$\log_a := f_a^{-1} : (0,+\infty) \to \mathbb{R}.$$

is called the logarithm to base a. *$\log_a$ is continuous and satisfies*

$$\log_a(x) = \frac{\log x}{\log a} \qquad \forall x \in (0,+\infty).$$

To see the last equality, consider $f_a\left(\frac{\log x}{\log a}\right) = \exp\left(\frac{\log x}{\log a}\cdot\log a\right) = \exp(\log(x)) = x$. Since f_a is bijective, this shows the desired equation.

Remark 3.73. (1) The formula from Theorem/Definition 3.72 shows that any logarithm differs only by multiplication with a constant value from the natural logarithm. In particular, properties b) and c) from Theorem 3.69 hold for logarithms to arbitrary bases, while $\log_a 1 = 0$ and $\log_a a = 1$ for all $a > 0$.

(2) In physics, the logarithm to base 10 plays an important role. If two positive numbers, e.g. measured values of an experiment, a and b are compared, then

$$\log_{10}\left(\frac{a}{b}\right)$$

is called the *order of magnitude* that a is bigger (or smaller) than b.

We conclude this section by discussing certain limits of logarithms and powers.

Theorem 3.74. *a)* $\lim\limits_{x\searrow 0}\log x = -\infty$, $\lim\limits_{x\to+\infty}\log x = +\infty$.

b) For all $s > 0$ it holds that

$$\lim_{x\searrow 0} x^s = 0. \qquad \lim_{x\to+\infty} x^s = +\infty.$$

c) For all $s \in (0, +\infty)$ it holds that

$$\lim_{x \to +\infty} \frac{\log x}{x^s} = 0.$$

Proof. a) By construction, $\log : (0, +\infty) \to \mathbb{R}$ is bijective. By Theorem 3.23 and the monotonicity of exp, log is strictly monotonically increasing. From these two properties, one easily derives the claim.

b) Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in $(0, +\infty)$ with $\lim_{n \to \infty} x_n = 0$. By a), $\lim_{n \to \infty} s \log(x_n) = -\infty$, so by Theorem 3.65,

$$\lim_{n \to \infty} x_n^s = \lim_{n \to \infty} \exp(s \log x_n) = 0.$$

Thus, it follows from Theorem 3.49 that $\lim_{x \searrow 0} x^s = 0$. The other limit is computed analogously.

c) Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in $(0, +\infty)$ diverging to $+\infty$. By a), the sequence $(s \log x_n)_{n \in \mathbb{N}}$ diverges to $+\infty$ as well. Since for each $n \in \mathbb{N}$ it holds that

$$\frac{\log x_n}{x_n^s} = \frac{1}{s} \frac{s \log x_n}{\exp(s \log x_n)},$$

it follows from Theorem 3.49 that

$$\lim_{n \to \infty} \frac{\log x_n}{x_n^s} = \frac{1}{s} \lim_{y \to +\infty} \frac{y}{e^y} = 0,$$

where the last limit is seen as in the proof of Theorem 3.65.

□

Remark 3.75. (1) Motivated by Theorem 3.74.b), we put $0^s := 0$ for all $s \in (0, +\infty)$.

(2) Loosely speaking, Theorem 3.74 says that for x tending to $+\infty$, $\log x$ grows slower to $+\infty$ than any power x^s, $s \in \mathbb{R}$. In particular, $\log x$ grows slower than $\sqrt[n]{x}$ as $x \to +\infty$ for each $n \in \mathbb{N}$.

3.6 Trigonometric functions

In high school one encounters the sine and the cosine of an angle in Euclidean geometry when studying right-angled triangles. We recall that for an angle α in a right-angled triangle, sine and cosine are given by

$$\sin \alpha = \frac{\text{length of opposite}}{\text{length of hypotenuse}}, \qquad \cos \alpha = \frac{\text{length of adjacent}}{\text{length of hypotenuse}}.$$

In this lecture, we employ an entirely different definition of sine and cosine which is well-defined within our rigid formal framework and which a priori looks very awkward and its connection to triangles is not at all obvious.

Definition 3.76. We define the *cosine* function by

$$\cos : \mathbb{R} \to \mathbb{R}, \qquad \cos(x) = \mathrm{Re}(e^{ix}),$$

and the *sine* function by

$$\sin \mathbb{R} \to \mathbb{R}, \qquad \sin(x) = \mathrm{Im}(e^{ix}).$$

Note that by definition *Euler's formula* is satisfied:

$$e^{ix} = \cos x + i \sin x \qquad \forall x \in \mathbb{R}.$$

How is this connected to the geometric definition of sine and cosine of an angle? We have seen in Theorem 2.99.g) that $|e^{ix}| = |\exp(ix)| = 1$ for each $x \in \mathbb{R}$. Geometrically, viewing $\mathbb{C}$ as a Euclidean plane, e^{ix} lies on the unit circle, i.e. the circle of radius 1 around 0. Assume that $x \in (0, \frac{\pi}{2})$. In this case, one shows that e^{ix} lies in the first quadrant (i.e. in $\{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0,\ \mathrm{Im}(z) > 0\}$. Construct a triangle as follows:

- draw the line segment connecting 0 and e^{ix},
- draw a vertical line segment from e^{ix} to the x-axis (i.e. the $\mathrm{Re}(z)$-axis),
- draw a horizontal line along the x-axis from 0 to the connection point of the vertical segment from before with the y-axis.

One realizes that the three line segments indeed form a right-angled triangle, whose hypotenuse is the segment connecting 0 with e^{ix}, which has length 1. Hence, the "geometric" sine and the cosine of the angle at 0 are given by the length of the opposite, which is $\mathrm{Im}(e^{ix}) = \sin x$, and the length of the adjacent, which is $\mathrm{Re}(e^{ix}) = \cos x$, respectively. Thus, the angle of the triangle at 0 is indeed given by x (measured as radian)! In other words, e^{ix} is the point on the unit circle whose connecting line to 0 forms the angle x with the horizontal axis. *(See the lecture videos for a drawing.)*

To get the connection to the sine and cosine for arbitrary right-angled triangles, one needs to show that angles are invariant under shifts, rotations and rescalings of the triangle. Then one easily sees that every triangle can be transformed into a triangle of the above form and defines sine and cosine as the sine and cosine of the transformed triangle to re-obtain the geometric definition. We will leave the details and proofs to the geometers and go on by studying the sine and cosine and their properties assuming our new definition.

In the following, we will occasionally use the shorthand notation

$$\sin^n(x) := (\sin x)^n \qquad \cos^n(x) := (\cos x)^n \qquad \forall n \in \mathbb{N}.$$

The next theorem summarizes several important elementary properties of sine and cosine.

Theorem 3.77. *a)* $\sin : \mathbb{R} \to \mathbb{R}$ *and* $\cos : \mathbb{R} \to \mathbb{R}$ *are continuous.*

b) $\cos x = \dfrac{e^{ix} + e^{-ix}}{2}, \qquad \sin x = \dfrac{e^{ix} - e^{-ix}}{2i} \quad \forall x \in \mathbb{R}.$

c) $\cos^2(x) + \sin^2(x) = 1$ *for all* $x \in \mathbb{R}$.

d) $|\sin x| \leq 1$ *and* $|\cos x| \leq 1$ *for all* $x \in \mathbb{R}$.

e) $\cos(-x) = \cos x$ *and* $\sin(-x) = -\sin x$ *for all* $x \in \mathbb{R}$.

Proof. a) This follows from the fact that exp is continuous and from Theorem 3.5.d).

b) This follows from Theorem 1.70.b), since $\overline{e^{ix}} = e^{\overline{ix}} = e^{-ix}$ by Theorem 2.99.f).

c) By Theorem 2.99.g), it holds for all $x \in \mathbb{R}$ that

$$1 = |e^{ix}|^2 = (\mathrm{Re}(e^{ix}))^2 + (\mathrm{Im}(e^{ix}))^2 = \sin(x)^2 + \cos(x)^2.$$

d) This follows directly from c).

e) This follows immediately from b).

□

Theorem 3.78 (Angle sum identities)**.** *Let $x, y \in \mathbb{R}$. Then*

$$\begin{aligned}\cos(x+y) &= \cos x \cos y - \sin x \sin y,\\ \sin(x+y) &= \sin x \cos y + \cos x \sin y.\end{aligned}$$

In particular,

$$\cos(2x) = \cos^2(x) - \sin^2(x) = 2\cos^2(x) - 1, \qquad \sin(2x) = 2\sin x \cos x.$$

Proof. Since $e^{i(x+y)} = e^{ix}e^{iy}$ by (3.2), it follows from Euler's formula that

$$\begin{aligned}\cos(x+y) + i\sin(x+y) &= (\cos x + i \sin x)(\cos y + i \sin y)\\ &= \cos x \cos y - \sin x \sin y + i(\sin x \cos y + \cos x \sin y)\end{aligned}$$

by the rules of complex multiplication. Considering real and imaginary parts separately shows a) and b). The latter formulas follow by taking $x = y$ and using that $\sin^2(x) = 1 - \cos^2(x)$ by Theorem 3.77.c). □

We derive from their definitions via the exponential function, which is a power series, that sine and cosine are described by power series as well.

Theorem 3.79. *For each $x \in \mathbb{R}$ it holds that*

$$\begin{aligned}\cos x &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots\\ \sin x &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots\end{aligned}$$

These series converge absolutely for all $x \in \mathbb{R}$.

Proof. Let $x \in \mathbb{R}$. By definition, $e^{ix} = \exp(ix) = \sum_{n=0}^{\infty} \frac{i^n x^n}{n!}$. For each $k \in \mathbb{N}_0$ it holds that $i^{2k} = (i^2)^k = (-1)^k$ and $i^{2k+1} = i(-1)^k$. Hence, we can split e^{ix} into its real and its imaginary part as follows:

$$e^{ix} = \sum_{k=0}^{\infty} \frac{i^{2k}x^{2k}}{(2k)!} + i\sum_{k=0}^{\infty} \frac{i^{2k}x^{2k+1}}{(2k+1)!} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} + i\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}.$$

This shows the claim by definition of cos and sin. The absolute convergence follows from the absolute convergence of $\exp(ix)$. □

Theorem 3.80. *For each $n \in \mathbb{N}_0$ and $x \in \mathbb{R}$ it holds that*

$$\cos x = \sum_{k=0}^{n} (-1)^k \frac{x^{2k}}{(2k)!} + r_{2n+2}(x), \qquad \sin x = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!} + r_{2n+3}(x),$$

with

$$|r_m(x)| \leq \frac{|x|^m}{m!} \qquad \text{if } |x| \leq m+1$$

for each $m \in \mathbb{N}$ with $m \geq 2$.

Proof. (This proof is not discussed in the lecture videos.)

We will only discuss the case of m even, the case of m odd follows analogously. By definition,

$$r_{2n+2}(x) = \sum_{k=n+1}^{\infty} (-1)^k \frac{x^{2k}}{(2k!)} = (-1)^{n+1} \frac{x^{2n+2}}{(2n+2)!} \sum_{k=0}^{\infty} (-1)^k a_k,$$

where $a_0 := 1$ and

$$a_k := \frac{x^{2k}}{(2n+3)(2n+4)\dots(2n+2(k+1))} \qquad \forall k \in \mathbb{N}.$$

Then for all $k \in \mathbb{N}$,

$$a_k = a_{k-1} \cdot \frac{x^2}{(2n+2k+1)(2n+2k+2)}.$$

This shows that if $|z| \leq 2n+3$, then $(a_n)_{n\in\mathbb{N}}$ is monotonically decreasing. Similar to the proof of the Leibniz test (Theorem 2.72), one derives that $0 \leq \sum_{k=0}^{\infty}(-1)^k a_k \leq 1$ and the claim follows by taking absolut values. □

The previous estimate enables us to compute an important limit.

Corollary 3.81. $\lim\limits_{x\to 0} \dfrac{\sin x}{x} = 1.$

Proof. In the notation of Theorem 3.79, it holds for each $x \neq 0$ that

$$\frac{\sin x}{x} = \frac{x + r_3(x)}{x} = 1 + \frac{r_3(x)}{x}.$$

Thus, by Theorem 3.79 it holds for all $x \in (-4, 0) \cup (0, 4)$ that

$$0 \leq \left| \frac{\sin x}{x} - 1 \right| = \frac{|r_3(x)|}{|x|} \leq \frac{|x|^2}{6}.$$

Since $\lim_{x\to 0} \frac{|x|^2}{6} = 0$, it follows from Theorem 3.60 that $\lim_{x\to 0} |\frac{\sin x}{x} - 1| = 0$, which shows the claim. □

We want to investigate the behaviour of sine and cosine as functions and study their properties.

Theorem 3.82. *a)* $\cos 0 = 1$ *and* $\sin 0 = 0$.

b) $\sin x > 0$ *for all* $x \in (0, 2]$.

c) $\cos|_{[0,2]}$ *is strictly monotonically decreasing and there is a unique* $x_0 \in (0,2)$ *with* $\cos(x_0) = 0$.

Proof. a) This is obvious, since $e^{i0} = e^0 = 1$.

b) We have seen in Theorem 3.80 that $\sin(x) = x + r_3(x)$, where $|r_3(x)| \leq \frac{|x|^3}{6}$ if $|x| \leq 4$. In particular, $r_3(x) \geq -\frac{x^3}{6}$ and thus for $x \in (0,2]$

$$\sin x \geq x - \frac{x^3}{6} = x\Big(1 - \frac{x^2}{6}\Big) \geq x\Big(1 - \frac{4}{6}\Big) = \frac{x}{3} > 0.$$

c) Let $x, y \in [0,2]$ with $x < y$ and put $u := \frac{x+y}{2}$, $v := \frac{y-x}{2}$. Then by Theorems 3.78 and 3.77

$$\begin{aligned}\cos x - \cos y &= \cos(u-v) - \cos(u+v)\\ &= \cos u\cos(-v) - \sin u\sin(-v) - (\cos u\cos v - \sin u\sin v)\\ &= \cos u\cos v + \sin u\sin(v) - \cos u\cos v + \sin u\sin v\\ &= 2\sin u\sin v = 2\sin\Big(\underbrace{\frac{x+y}{2}}_{\in(0,2]}\Big)\sin\Big(\underbrace{\frac{y-x}{2}}_{\in[0,2]}\Big) > 0,\end{aligned}$$

where the last inequality follows from a). Hence $\cos|_{[0,2]}$ is strictly monotonically decreasing. In particular, cos is injective on $[0,2]$. Since $\cos 0 = 1$, the existence of a unqiue $x_0 \in (0,2)$ with $\cos(x_0) = 0$ follows from the intermediate value theorem if we can show that $\cos 2 \overset{!}{<} 0$. By Theorem 3.80, it holds for $x \in [-5,5]$ that

$$\cos x = 1 - \frac{x^2}{2} + r_4(x), \quad \text{where } |r_4(x)| \leq \frac{|x|^4}{24}.$$

This shows for $x \in [-5,5]$ that $\cos(x) \leq 1 - \frac{x^2}{2} + \frac{x^4}{24}$ and in particular

$$\cos 2 \leq 1 - \frac{4}{2} + \frac{16}{24} = -1 + \frac{2}{4} = -\frac{1}{3} < 0.$$

This completes the proof. □

While the definition of sine of cosine may already have been awkward at first sight, the following definition of the number π might look even more awkward to many of you.

Definition 3.83. We define $\pi \in \mathbb{R}$ by

$$\pi := 2 \cdot \min\{x \in (0,+\infty) \mid \cos x = 0\}.$$

By Theorem 3.82.b) π is a well-defined real number with $0 < \pi < 4$.
In your high school course on geometry, π was probably introduced in terms of circles. More precisely, given an arbitrary circle in the Euclidean plane $\mathbb{R}^2$, π is given by

$$\pi = \frac{\text{circumference of the circle}}{\text{diameter of the circle}}.$$

Of course one needs to show that this ratio is indeed the same for every circle. Here, it is indeed possible to prove that our definition of π coincides with the geometric one, but we shall again leave the details to the geometers.
Instead, we shall compute several values of sine and cosine related to π.

Theorem 3.84.

$$e^{i\frac{\pi}{2}} = i, \qquad e^{i\pi} = -1, \qquad e^{i\frac{3}{2}\pi} = -i, \qquad e^{i2\pi} = 1,$$

which implies the following values for sine and cosine:

x	0	$\frac{\pi}{2}$	π	$\frac{3}{2}\pi$	2π
$\sin x$	0	1	0	-1	0
$\cos x$	1	0	-1	0	1

Proof. Since $\cos(\frac{\pi}{2}) = 0$ by construction, it follows from Theorem 3.77.c) that $\sin^2(\frac{\pi}{2}) = 1 - \cos^2(\frac{\pi}{2}) = 1$. By Theorem 3.82.b) this yields $\sin(\frac{\pi}{2}) = 1$ and thus

$$e^{i\frac{\pi}{2}} = \cos\left(\tfrac{\pi}{2}\right) + i\sin\left(\tfrac{\pi}{2}\right) = i.$$

Consequently,

$$e^{i\pi} = \left(e^{i\frac{\pi}{2}}\right)^2 = i^2 = -1, \qquad e^{i\frac{3}{2}\pi} = \left(e^{i\frac{\pi}{2}}\right)^3 = i^3 = -i, \qquad e^{i2\pi} = \left(e^{i\frac{\pi}{2}}\right)^4 = i^4 = 1.$$

□

Remark 3.85. As part of the previous theorem, we obtain *Euler's identity*,

$$e^{i\pi} + 1 = 0.$$

In 1990 this equation was voted "the most beautiful theorem in mathematics" by readers of The Mathematical Intelligencer. It unites the most important numbers in mathematics, 0, 1, e, i and π, in one concise formula and connects them by addition, multiplication and the exponential function. Sheer beauty, don't you think?

From the previous properties and equation (3.2) for the exponential function we derive periodicity properties for sine and cosine.

Corollary 3.86. *Let $x \in \mathbb{R}$. Then:*

a) $\cos(x + 2\pi) = \cos x, \quad \sin(x + 2\pi) = \sin x.$

b) $\cos(x + \pi) = -\cos x, \quad \sin(x + \pi) = -\sin x.$

c) $\cos x = \sin(\frac{\pi}{2} - x), \quad \sin x = \cos(\frac{\pi}{2} - x).$

Proof. In all three parts, we apply Theorem 3.84 and the equality (3.2).

a) For all $x \in \mathbb{R}$, $e^{i(x+2\pi)} = e^{ix+i2\pi} = e^{ix}e^{i2\pi} = e^{ix}$. Taking real and imaginary parts shows the claim.

b) For all $x \in \mathbb{R}$, $e^{i(x+\pi)} = e^{ix+i\pi} = e^{ix}e^{i\pi} = -e^{ix}$. Thus,

$$\cos(x + \pi) = \mathrm{Re}(-e^{ix}) = -\mathrm{Re}(e^{ix}) = -\cos x,$$
$$\sin(x + \pi) = \mathrm{Im}(-e^{ix}) = -\mathrm{Im}(e^{ix}) = -\sin x.$$

c) For all $x \in \mathbb{R}$, $e^{i(\frac{\pi}{2}-x)} = e^{i\frac{\pi}{2}}e^{-ix} = ie^{-ix} =$. By Theorem 3.77.e),

$$ie^{-ix} = i(\cos(-x) + i\sin(-x)) = i(\cos x - i\sin x) = \sin x + i\cos x.$$

Consequently,

$$\cos\left(\tfrac{\pi}{2} - x\right) = \mathrm{Re}(ie^{-ix}) = \sin x, \qquad \sin\left(\tfrac{\pi}{2} - x\right) = \mathrm{Im}(ie^{-ix}) = \cos x.$$

□

Corollary 3.87. *Let $x \in \mathbb{R}$. Then:*

a) $\sin x = 0$ *if and only if* $x \in \{k\pi \mid k \in \mathbb{Z}\}$.

b) $\cos x = 0$ *if and only if* $x \in \{\frac{2k+1}{2}\pi \mid k \in \mathbb{Z}\}$.

Proof. a) Since $\sin 0 = 0$, it follows from Corollary 3.86.a) that $\sin(k\pi) = 0$ for all $k \in \mathbb{Z}$. We know from Theorem 3.82 that $\sin x > 0$ for $x \in (0, \frac{\pi}{2}]$. We derive from Theorem 3.82 and Theorem 3.77.e) that $\cos x > 0$ for all $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$. From this and from Corollary 3.86.c) that

$$\sin(x) = \cos\left(\tfrac{\pi}{2} - x\right) > 0 \qquad \forall x \in \left(\tfrac{\pi}{2}, \pi\right).$$

Thus, $\sin(x) \neq 0$ for all $x \in (0, \pi)$. By Corollary 3.86.b), this implies $\sin(x) \neq 0$ for all $x \in (\pi, 2\pi)$ and by iterating this argument we obtain $\sin x \neq 0$ for all $x \in (k\pi, (k+1)\pi)$, $k \in \mathbb{Z}$. Together with the above, this shows the claim.

b) This follows directly from a) and Corollary 3.86.c).

□

Using the previous information, we can draw graphs of the qualitative behaviour of sine and cosine. See the lecture videos for such an illustration.

Corollary 3.88. *For $x \in \mathbb{R}$ it holds that $e^{ix} = 1$ if and only if $x = 2k\pi$ for some $k \in \mathbb{Z}$.*

Proof. By Theorem 3.77.b),

$$\sin\left(\frac{x}{2}\right) = \frac{1}{2i}(e^{i\frac{x}{2}} - e^{-i\frac{x}{2}}) = \frac{e^{-i\frac{x}{2}}}{2i}(e^{ix} - 1).$$

Since $\frac{e^{-i\frac{x}{2}}}{2i} \neq 0$ for each x it follows that $e^{ix} = 1$ if and only if $\sin(\frac{x}{2}) = 0$. The claim thus follows from Corollary 3.87. □

Another important function involving sine and cosine that we will not discuss in detail at this point is the following.

Definition 3.89. The *tangent function* is the function

$$\tan : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}, \qquad \tan x := \frac{\sin x}{\cos x}.$$

$\tan x$ is well-defined since, as we already noted in the proof of Corollary 3.87, $\cos x > 0$ for all $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

Theorem/Definition 3.90. *a)* $\cos|_{[0,\pi]}$ *is strictly monotonically decreasing and maps* $[0,\pi]$ *bijectively onto* $[-1,1]$. *The inverse function*

$$\arccos := (\cos|_{[0,\pi]})^{-1} : [-1,1] \to [0,\pi]$$

is continuous and is called the arccosine.

b) $\sin|_{[-\frac{\pi}{2},\frac{\pi}{2}]}$ *is strictly monotonically decreasing and maps* $[-\frac{\pi}{2},\frac{\pi}{2}]$ *bijectively onto* $[-1,1]$. *The inverse function*

$$\arcsin := (\sin|_{[-\frac{\pi}{2},\frac{\pi}{2}]})^{-1} : [-1,1] \to \Big(-\frac{\pi}{2},\frac{\pi}{2}\Big)$$

is continuous and is called the arcsine.

c) $\tan : (-\frac{\pi}{2},\frac{\pi}{2}) \to \mathbb{R}$ *is continuous, strictly monotonically increasing and bijective. Its inverse function*

$$\arctan := \tan^{-1} : \mathbb{R} \to \Big(-\frac{\pi}{2},\frac{\pi}{2}\Big)$$

is continuous and is called the arctangent.

Proof. (This proof is not discussed in the lecture videos.)

a) By Theorem 3.82, $\cos|_{[0,\frac{\pi}{2}]}$ is strictly monotonically decreasing. Thus, for $x,y \in [\frac{\pi}{2},\pi]$ with $x < y$ it holds that $\cos(\pi - y) > \cos(\pi - x)$ and thus by Theorem 3.77.e) and Corollary 3.86.b),

$$\cos(x) = \cos(-x) = -\cos(\pi - x) > -\cos(\pi - y) = \cos(y).$$

It follows that $\cos|_{[0,\pi]}$ is stricly monotonically decreasing. Since $\cos(0) = 1$ and $\cos(\pi) = -1$, the rest of the claim follows from Theorem 3.23.

b) Since $\sin(x) = \cos(\frac{\pi}{2} - x)$ for all $x \in \mathbb{R}$, $\sin(-\frac{\pi}{2}) = -1$, $\sin(\frac{\pi}{2}) = 1$, this follows from part a).

c) The continuity of tan follows from the continuity of sin and cos. Let $x,y \in [0,\frac{\pi}{2})$. Then as seen in a) and b), $\sin x < \sin y$ and $\cos x > \cos y$, such that

$$\tan x = \frac{\sin x}{\cos x} < \frac{\sin y}{\cos y} = \tan(y).$$

Hence $\tan|_{[0,\frac{\pi}{2})}$ is strictly monotonically increasing. Since

$$\tan(x) = \frac{\sin x}{\cos x} = \frac{-\sin(-x)}{\cos(-x)} = -\tan(-x),$$

one derives that $\tan : (-\frac{\pi}{2},\frac{\pi}{2}) \to \mathbb{R}$ is strictly monotonically increasing, hence injective.

Claim: $\lim\limits_{x\to\frac{\pi}{2}} \tan x = +\infty$ and $\lim\limits_{x\to-\frac{\pi}{2}} \tan x = -\infty$.

Proof of Claim. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $(-\frac{\pi}{2},\frac{\pi}{2})$ with $\lim_{n\to\infty} x_n = \frac{\pi}{2}$. Assume w.l.o.g. that $x_n > 0$ for all $n \in \mathbb{N}$ and put

$$y_n := \frac{\cos x_n}{\sin x_n}.$$

By Theorem 3.82, $y_n > 0$ for all $n \in \mathbb{N}$ and by the continuity of sine and cosine, it holds that

$$\lim_{n\to\infty} y_n = \frac{\lim_{n\to\infty} \cos x_n}{\lim_{n\to\infty} \sin x_n} = \frac{\cos(\frac{\pi}{2})}{\sin(\frac{\pi}{2})} = 0.$$

Hence, by Theorem 2.43, $\lim_{n\to\infty} \tan x_n = \lim_{n\to\infty} \frac{1}{y_n} = +\infty$. Since the sequence was arbitrary, $\lim_{x\to\frac{\pi}{2}} \tan x = +\infty$. Using $\tan(-x) = -\tan x$, one further derives obtains that $\lim_{x\to-\frac{\pi}{2}} \tan x = -\infty$. ■

From this claim and Theorem 3.23 one obtains that tan is surjective as well.

□

See the lecture videos for sketches of the graphs of the functions from Theorem/Definition 3.90.

In the end of this section, we want to use the exponential function to derive a different way of displaying complex numbers than their usual description by real and imaginary parts.

Theorem 3.91 (Polar coordinates)**.** *For every complex number $z \in \mathbb{C}$ there are numbers $r \in [0, +\infty)$ and $\varphi \in \mathbb{R}$, such that*

$$z = re^{i\varphi}.$$

If $r_1 \neq 0$, then

$$r_1 e^{i\varphi_1} = r_2 e^{i\varphi_2} \quad \Leftrightarrow \quad r_1 = r_2 \ \wedge \ \exists k \in \mathbb{Z} \text{ with } \varphi_1 = \varphi_2 + 2k\pi.$$

Remark 3.92. (1) For $z = re^{i\varphi}$ it holds that $|z| = r|e^{i\varphi}| = r$. This shows that r is the length of the line segment from the origin to z.

(2) By the explanations on sine and cosine above, φ is the the angle between the line segment from 0 to z and the x-axis. (See the illustration in the lecture videos.) φ is also called the *argument of z*.

(3) These observations yield a geometric interpretation of complex multiplication. If $z_1 = r_1 e^{i\varphi_1}$ and $z_2 = r_2 e^{i\varphi_2}$, then

$$z_1 \cdot z_2 = r_1 r_2 e^{i\varphi_1} e^{i\varphi_2} = r_1 r_2 e^{i(\varphi_1+\varphi_2)}.$$

Thus, $z_1 \cdot z_2$ is the complex number, whose absolute value is the product of the absolute values of z_1 and z_2 and whose angle to the x-axis is the sum of the angles of z_1 and z_2. (See the illustration in the lecture.)

Proof of Theorem 3.91. If $z = 0$, then $z = 0 \cdot e^{i\varphi}$ for all $\varphi \in \mathbb{R}$. Assume $z \neq 0$ and put $r := |z|$ and $w := \frac{z}{r}$, such that $|w| = 1$. Let $a := \mathrm{Re}(w)$ and $b := \mathrm{Im}(w)$. Then

$$a^2 + b^2 = |w|^2 = 1, \tag{3.3}$$

such that $|a| \leq 1$. Hence, $\alpha := \arccos(a) \in [0, \pi]$ is well-defined. Then $\cos\alpha = a$, such that by (3.3), $|b| = |\sin\alpha|$. Put

$$\varphi := \begin{cases} \alpha & \text{if } b = \sin\alpha, \\ -\alpha & \text{if } b = -\sin\alpha. \end{cases}$$

Then $\cos\varphi = \cos\alpha = a$ and $\sin\varphi = b$, such that, by Euler's formula, $w = e^{i\varphi}$ and thus $z = re^{i\varphi}$.

Let $r_1 e^{i\varphi_1} = r_2 e^{i\varphi_2}$ with $r_1 \neq 0$. Then

$$\frac{r_2}{r_1} = \frac{e^{i\varphi_1}}{e^{i\varphi_2}} = e^{i(\varphi_1 - \varphi_2)}.$$

Since $|e^{i(\varphi_1-\varphi_2)}| = 1$, this yields $r_1 = r_2$ and $e^{i(\varphi_1-\varphi_2)} = 1$. The claim thus follows from Corollary 3.88. □

From Euler's formula and from the proof of Theorem 3.91 one can directly derive for the representation $z = x + iy$ and representations of the form $z = re^{i\varphi}$ how one gets from one to the other:

- If $z = re^{i\varphi}$ is given, then $x = r\cos\varphi$ and $y = r\sin\varphi$.
- If $z = x + iy$ is given, then $r = |z| = \sqrt{x^2 + y^2}$ and $\varphi = \begin{cases} \arccos(\frac{x}{r}) & \text{if } y \geq 0, \\ -\arccos(\frac{x}{r}) & \text{if } y < 0. \end{cases}$

We conclude this section by deriving a result about complex solutions of a simple, but important equation.

Corollary 3.93. *For each $n \in \mathbb{N}$ the equation*

$$z^n = 1$$

has precisely n distinct solutions $\{z_0, z_1, \ldots, z_{n-1}\}$ in $\mathbb{C}$. Explicitly,

$$z_k = e^{i\frac{2\pi k}{n}} \qquad \forall k \in \{0, 1, \ldots, n-1\}.$$

Proof. (This proof is not discussed in the lecture videos.)

Let $z \in \mathbb{C}$ with $z^n = 1$. By Theorem 3.91 there are $r \in (0, +\infty)$ and $\varphi \in \mathbb{R}$ with $z = re^{i\varphi}$. Since $z^n = 1$, it holds that $|z|^n = 1$, which implies $r = |z| = 1$, so $z = e^{i\varphi}$. Then

$$1 = (e^{i\varphi})^n = e^{in\varphi} = \cos(n\varphi) + i\sin(n\varphi).$$

In particular, $\sin(n\varphi) = 0$ and thus, by Corollary 3.87 there exists $m \in \mathbb{N}$ with $n\varphi = m\pi$. Since it also needs to hold that $\cos(n\varphi) = 1$ and since by Theorems 3.84 and 3.84

$$\cos(m\pi) = \begin{cases} 1 & \text{if } m \text{ is even}, \\ -1 & \text{if } m \text{ is odd}, \end{cases}$$

it follows that m must be even. Hence $e^{in\varphi} = 0$ if and only if there exists $k \in \mathbb{Z}$ with

$$n\varphi = 2k\pi \qquad \Leftrightarrow \qquad \varphi = \frac{2k\pi}{n}.$$

Put $z_k = e^{i\frac{2k\pi}{n}}$ for each $k \in \mathbb{Z}$. It follows from Theorem 3.84 that

$$z_{k+n} = e^{i\frac{2(k+n)\pi}{n}} = e^{i(\frac{2k\pi}{n} + 2\pi)} = e^{i\frac{2k\pi}{n}} e^{2\pi i} = e^{i\frac{2k\pi}{n}} = z_k$$

for each $k \in \mathbb{Z}$. This implies the claim. □

Chapter 4

Differentiation

4.1 Differentiability and derivatives

In this section, we will finally introduce the notion of differentiability which many of you might have expected to see much earlier in this course and that you have already used in the physics classes. We shall see that its formal definition indeed requires limits of functions, so we will use the formalism from the previous chapter.

Definition 4.1. Let I be an interval, $x_0 \in I$ and $f : I \to \mathbb{C}$. f is *differentiable in* x_0 if the limit

$$f'(x_0) := \lim_{h \to 0} \frac{f(x_0 + h) - f(x)}{h}$$

exists. $f'(x_0)$ is then called the *derivative of* f *in* x_0 or the *differential quotient of* f *in* x_0. We call f *differentiable* if it is differentiable in every $x_0 \in I$.

Remark 4.2. (1) The differential quotient is also characterized by

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which many authors use to define $f'(x_0)$.

(2) *Geometric interpretation of the differential quotient:*
Let I be an interval, $f : I \to \mathbb{R}$ and $x_0 \in I$. Given $h \neq 0$, the number

$$m(x_0; h) := \frac{f(x_0 + h) - f(x_0)}{h}$$

is the slope of the line in $\mathbb{R}^2$ going through $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$. Such a line is called a *secant line* of the graph of f. $f'(x_0) = \lim_{h \to 0} m(x_0; h)$ is thus given as the limit of the slopes of the secant lines going through $(x_0, f(x_0))$ as the second point wanders along the graph of f towards $(x_0, f(x_0))$. (See the illustration in the lecture videos.) This yields that $f'(x_0)$ *is the slope of the tangent line to the graph of* f *at* x_0, i.e. the line which most closely approximates the graph of f at x_0. This will be made more precise in Theorem 4.6 below.

(3) Let I be an interval, $u, v : I \to \mathbb{R}$ and let $f : I \to \mathbb{C}$, $f(x) = u(x) + iv(x)$. Then f is differentiable in $x_0 \in I$ if and only if both u and v are differentiable in x_0. It then holds that

$$f'(x_0) = u'(x_0) + iv'(x_0).$$

This is easily seen since splitting the differential quotient into real and imaginary part.

The following theorem contains computations of the derivatives of the most elementary functions that we have seen in this course.

Theorem 4.3. *a) A constant function $f : \mathbb{R} \to \mathbb{C}$, $f(x) = c$ for all $x \in \mathbb{R}$, where $c \in \mathbb{C}$, is differentiable with $f'(x) = 0$ for all $x \in \mathbb{R}$.*

b) Let $n \in \mathbb{N}$. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^n$, is differentiable with

$$f'(x) = nx^{n-1} \qquad \forall x \in \mathbb{R}.$$

c) $f : (0, +\infty) \to \mathbb{R}$, $f(x) = \frac{1}{x}$, is differentiable with

$$f'(x) = -\frac{1}{x^2} \qquad \forall x \in (0, +\infty).$$

d) Let $c \in \mathbb{C}$. Then $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{cx}$, is differentiable with

$$f'(x) = ce^{cx} \qquad \forall x \in \mathbb{R}.$$

Proof. a) For each $x \in \mathbb{R}$ we compute that

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{c - c}{h} = \lim_{h\to 0} 0 = 0,$$

which shows the claim.

b) For all $x, h \in \mathbb{R}$ with $h \neq 0$ we use the binomial theorem to compute:

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^n - x^2}{h} = \frac{1}{h}\Big(\sum_{k=0}^{n} \binom{n}{k} x^{n-k}h^k - x^n\Big) = \frac{1}{h}\sum_{k=1}^{n} \binom{n}{k} x^{n-k}h^k$$
$$= \sum_{k=1}^{n} \binom{n}{k} x^{n-k}h^{k-1} = \sum_{k=0}^{n-1} \binom{n}{k+1} x^{n-1-k}h^k,$$

where we performed an index shift in the final step. Hence,

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \sum_{k=0}^{n-1} \binom{n}{k+1} x^{n-1-k}h^k = \binom{n}{1} x^{n-1} = nx^{n-1}.$$

c) For all $x, h \in \mathbb{R}$ with $h \neq 0$ we compute:

$$\frac{f(x+h) - f(x)}{h} = \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \frac{x - (x+h)}{hx(x+h)} = -\frac{1}{x(x+h)}.$$

Thus,

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \Big(-\frac{1}{x(x+h)}\Big) = -\frac{1}{x^2}.$$

d) We first compute that for $h \neq 0$:

$$\frac{e^{ch}-1}{h} = \frac{1}{h}\Big(\sum_{n=0}^{\infty}\frac{c^n h^n}{n!} - 1\Big) = \frac{1}{h}\sum_{n=1}^{\infty}\frac{c^n h^n}{n!} = \sum_{n=1}^{\infty}\frac{c^n h^{n-1}}{n!} = \sum_{n=0}^{\infty}\frac{c^{n+1}h^n}{(n+1)!},$$

where we performed an index shift in the final step. By Corollary 3.43, i.e. by the continuity of power series, this yields

$$\lim_{h\to 0}\frac{e^{ch}-1}{h} = \lim_{h\to 0}\sum_{n=0}^{\infty}\frac{c^{n+1}h^n}{(n+1)!} = \sum_{n=0}^{\infty}\frac{c^{n+1}0^n}{(n+1)!} = \frac{c}{1!} = c.$$

For each $x \in \mathbb{R}$ we thus derive:

$$f'(x) = \lim_{h\to 0}\frac{e^{c(x+h)}-e^{cx}}{h} = \lim_{h\to 0}\frac{e^{cx}e^{ch}-e^{cx}}{h} = \lim_{h\to 0} e^{cx}\frac{e^{ch}-1}{h} = e^{cx}\lim_{h\to 0}\frac{e^{ch}-1}{h} = ce^{cx},$$

which shows the claim.

□

Corollary 4.4. *a)* $\exp : \mathbb{R} \to \mathbb{R}$ *is differentiable with* $\exp'(x) = \exp(x)$ *for all* $x \in \mathbb{R}$.

b) Let $a \in (0, +\infty)$. *Then* $f : \mathbb{R} \to \mathbb{R}$, $f(x) = a^x$, *is differentiable with*

$$f'(x) = (\log a)a^x \qquad \forall x \in \mathbb{R}.$$

c) $\sin, \cos : \mathbb{R} \to \mathbb{R}$ *are differentiable with*

$$\sin'(x) = \cos x, \qquad \cos'(x) = -\sin x \qquad \forall x \in \mathbb{R}.$$

Proof. a) This is Theorem 4.3.d) with $c = 1$.

b) This follows immediately from Theorem 4.3.d) with $c = \log a$, since $a^x = e^{\log a \cdot x}$ for all $x \in \mathbb{R}$.

c) By Theorem 4.3.d), the function $f : \mathbb{R} \to \mathbb{C}$, $f(x) = e^{ix} = \cos x + i \sin x$, satisfies $f'(x) = ie^{ix}$ for all $x \in \mathbb{R}$. By Remark 4.2.(3) this yields

$$\cos'(x) + i\sin'(x) = i(\cos x + i \sin x) = -\sin x + i \cos x.$$

Comparing real and imaginary parts shows the claim.

□

Example 4.5. The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$, is *not* differentiable in 0. Let $h_n = \frac{(-1)^n}{n}$ for each $n \in \mathbb{N}$. Then $\lim_{n\to\infty} h_n = 0$, but

$$\lim_{n\to\infty}\frac{f(h_n)-f(0)}{h_n} = \frac{|h_n|}{h_n} = \frac{\frac{1}{n}}{\frac{(-1)^n}{n}} = (-1)^n,$$

which is divergent. Hence, $\lim_{h\to 0}\frac{f(h)-f(0)}{h}$ does not exist.

Theorem 4.6. *Let* I *be an interval,* $f : I \to \mathbb{C}$ *and* $a \in I$. *The following are equivalent:*

(i) f *is differentiable in* a,

(ii) there are $c \in \mathbb{C}$ *and* $\varphi : I \to \mathbb{C}$ *with* $\lim_{x\to a} \frac{\varphi(x)}{x-a} = 0$, *such that*

$$f(x) = f(a) + c(x-a) + \varphi(x) \qquad \forall x \in I,$$

In this case, $c = f'(a)$.

(iii) there is a function $r : I \to \mathbb{C}$ *which is continuous in* a, *such that*

$$f(x) = f(a) + r(x)(x-a). \tag{4.1}$$

In this case, $r(a) = f'(a)$.

Proof. (i) $\Rightarrow$ (ii): Put $c := f'(a)$. and define $\varphi : I \to \mathbb{C}$, $\varphi(x) := f(x) - f(a) - f'(a)(x-a)$. Then

$$\lim_{x\to a} \frac{\varphi(x)}{x-a} = \lim_{x\to a} \left(\frac{f(x)-f(a)}{x-a} - f'(a) \right) = \lim_{x\to a} \frac{f(x)-f(a)}{x-a} - f'(a) = 0.$$

(ii) $\Rightarrow$ (iii): The function

$$r(x) = \begin{cases} c + \frac{\varphi(x)}{x-a} & \text{if } x \neq a, \\ c & \text{if } x = a, \end{cases}$$

has the desired properties.
(iii) $\Rightarrow$ (i): If such an r exists, then

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \lim_{x\to a} \frac{r(x)(x-a)}{x-a} = \lim_{x\to a} r(x) = r(a).$$

Hence, f is differentiable in a with $f'(a) = r(a)$. □

Remark 4.7. The key consequence of Theorem 4.6 is that *f is differentiable in x if and only if it can be locally approximated near x by an affine-linear function.*[1] Here, an affine-linear function $f : \mathbb{R} \to \mathbb{R}$ is a function of the form $f(x) = ax + b$, where $a, b \in \mathbb{R}$, i.e. a polynomial of degree one. Hence, to differentiate a function means to approximate a possibly very complicated function by a simple one. This approached will be further exploited in the course of this chapter.

Corollary 4.8. *Let I be an interval and $f : I \to \mathbb{C}$. If f is differentiable in $a \in I$, then f is continuous in a.*

Proof. This follows immediately from Theorem 4.6, since the right-hand side of (4.1) is continuous in a. □

As we have seen in Example 4.5, there are continuous functions which are not differentiable, hence the converse of Corollary 4.8 is false in general.

4.2 Rules of differentiation

In the previous sections we have computed the derivatives of certain functions explicitly by studying differential quotients. For more complicated functions, this is not a feasible method, since these differential quotients may not at all be easy to determine. In this section, we will discuss computational rules that allow us to compute derivatives of many important functions from our knowledge of derivatives of simple ones.

[1] This idea should be tattooed on your celebral cortex and is something that will be generalized to higher dimensions in next semester's Math lecture.

Theorem 4.9 (Algebraic operations and differentiation)**.** *Let I be an interval, $a \in I$, and let $f, g : I \to \mathbb{C}$ be differentiable in a. Then:*

a) (Linearity) *$f + g : I \to \mathbb{C}$ is differentiable in a with*

$$(f+g)'(a) = f'(a) + g'(a).$$

For each $\lambda \in \mathbb{C}$ the function $\lambda f : I \to \mathbb{C}$ is differentiable in a with

$$(\lambda f)'(a) = \lambda \cdot f'(a).$$

b) (Product rule) *$f \cdot g : I \to \mathbb{C}$ is differentiable in a with*

$$(f \cdot g)'(a) = f'(a) \cdot g(a) + f(a) \cdot g'(a).$$

c) (Quotient rule) *If $g(x) \neq 0$ for all $x \in I$, then $\frac{f}{g} : I \to \mathbb{C}$ is differentiable in a with*

$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{(g(a))^2}.$$

Proof. a) By Corollary 4.8, f and g are continuous in a. The claim thus follows directly from the corresponding properties of limits of functions from Theorem 3.50.

b) We compute that

$$\begin{aligned}
(f \cdot g)'(a) &= \lim_{h \to 0} \frac{f(a+h)g(a+h) - f(a)g(a)}{h} \\
&= \lim_{h \to 0} \frac{f(a+h)g(a+h) - f(a+h)g(a) + f(a+h)g(a) - f(a)g(a)}{h} \\
&= \lim_{h \to 0} \left(f(a+h)\frac{g(a+h) - g(a)}{h} + \frac{f(a+h) - f(a)}{h} g(a) \right) \\
&= \lim_{h \to 0} f(a+h) \cdot \lim_{h \to 0} \frac{g(a+h) - g(a)}{h} + \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} g(a) \\
&= f(a) \cdot g'(a) + f'(a) \cdot g(a).
\end{aligned}$$

Here, we have used that f is continuous in a by Corollary 4.8.

c) We first consider the case of $f(x) = 1$ for all $x \in I$. Then:

$$\begin{aligned}
\left(\frac{1}{g}\right)'(a) &= \lim_{h \to 0} \frac{\frac{1}{g(a+h)} - \frac{1}{g(a)}}{h} = \lim_{h \to 0} \frac{g(a) - g(a+h)}{h \cdot g(a+h) \cdot g(a)} \\
&= -\lim_{h \to 0} \frac{g(a+h) - g(a)}{h} \cdot \lim_{h \to 0} \frac{1}{g(a+h)g(a)} = \frac{-g'(a)}{(g(a))^2}.
\end{aligned}$$

Thus, by the product rule,

$$\begin{aligned}
\left(\frac{f}{g}\right)'(a) = \left(f \cdot \frac{1}{g}\right)'(a) &= f'(a)\frac{1}{g(a)} + f(a)\left(\frac{1}{g}\right)'(a) \\
&= \frac{f'(a)}{g(a)} - \frac{f(a)g'(a)}{(g(a))^2} = \frac{f'(a)g(a) - f(a)g'(a)}{(g(a))^2}.
\end{aligned}$$

□

Example 4.10. (1) Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sum_{k=0}^{n} a_k x^k$, where $a_0, a_1, \ldots, a_n \in \mathbb{R}$, be a real polynomial. Then by Theorem 4.9.a) and the computations from Theorem 4.3.a) and b), f is differentiable with

$$f'(x) = \sum_{k=1}^{n} a_k k x^{k-1} \qquad \forall x \in \mathbb{R}.$$

(2) By Theorem 4.3.b) it holds for $f : (0, +\infty) \to \mathbb{R}$, $f(x) = \frac{1}{x^n} = x^{-n}$, where $n \in \mathbb{N}$, by the quotient rule that

$$f'(x) = -\frac{nx^{n-1}}{(x^n)^2} = -\frac{nx^{n-1}}{x^{2n}} = -\frac{n}{x^{n+1}} = -nx^{-n-1}.$$

for each $x \in (0, +\infty)$.

(3) By the quotient rule and Corollary 4.4.c), we compute for each $x \in \mathbb{R}$ that

$$\tan'(x) = \left(\frac{\sin}{\cos}\right)'(x) = \frac{\cos x \cdot \cos x - \sin x \cdot (-\sin x)}{\cos^2(x)} = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)}.$$

Next we investigate the behaviour of differentiation with respect to the composition of functions.

Theorem 4.11 (Chain rule)**.** *Let I and J be intervals. Let $f : I \to \mathbb{R}$ and $g : J \to \mathbb{R}$ with $f(I) \subset J$. If f is differentiable in $a \in I$ and if g is differentiable in $f(a)$, then $g \circ f : I \to \mathbb{R}$ is differentiable in a with*

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

Proof. Since g is differentiable in $f(a)$, by Theorem 4.6 there exists a map $r : J \to \mathbb{C}$ with $r((f(a)) = g'(f(a))$ that is continuous in $f(a)$, such that

$$g(y) - g(f(a)) = r(y)(y - f(a)) \qquad \forall y \in J.$$

For $x \in I$ with $x \neq a$, this yields

$$\frac{g(f(x)) - g(f(a))}{x - a} = r(f(x)) \frac{f(x) - f(a)}{x - a}.$$

Since $r(f(a)) = g'(f(a))$ and since f is differentiable in a, we conclude

$$(g \circ f)'(a) = \lim_{x \to a} \frac{g(f(x)) - g(f(a))}{x - a} = \lim_{x \to a} r(f(x)) \cdot \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = g'(f(a)) \cdot f'(a).$$

□

Example 4.12. Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{-x^2}$. Then $f = g \circ h$, where $g : \mathbb{R} \to \mathbb{R}$, $g(x) = e^x$, and $h : \mathbb{R} \to \mathbb{R}$, $h(x) = -x^2$. Both of them are differentiable, hence f is differentiable by Theorem 4.11 and it holds that

$$f'(x) = g'(h(x)) \cdot h(x) = e^{-x^2}(-2x) = -2xe^{-x^2} \qquad \forall x \in \mathbb{R}.$$

We conclude this section by investigating the differentiability of inverse functions.

Theorem 4.13 (Derivatives of inverse functions)**.** *Let I be an interval consisting of more than one point and let $f : I \to \mathbb{R}$ be continuous and strictly monotonic. Put $J := f(I)$. If f is differentiable in $a \in I$ with $f'(a) \neq 0$, then its inverse function $f^{-1} : J \to I$ is differentiable in $b := f(a)$ with*

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

Proof. Since f is differentiable in a, by Theorem 4.6 there exists $r : I \to \mathbb{R}$ which is continuous in a, such that $r(a) = f'(a)$ and

$$f(x) = f(a) + r(x)(x - a) \qquad \forall x \in I.$$

Since $f'(a) \neq 0$ and since f is strictly monotonic, $r(x) \neq 0$ for all $x \in I$. Thus, the above equation is equivalent to

$$x - a = \frac{1}{r(x)}(f(x) - f(a)) \forall x \in I,$$

which yields

$$f^{-1}(y) - f^{-1}(b) = \frac{1}{r(f^{-1}(y))}(y - b) \qquad \forall y \in J.$$

By Theorem 3.23, f^{-1} is continuous in b which yields that $\frac{1}{r \circ f^{-1}}$ is continuous in b. Hence, by Theorem 4.6, it follows that f^{-1} is differentiable in b with

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

□

Example 4.14. (1) $\log : (0, +\infty) \to \mathbb{R}$ is the inverse function of $\exp : \mathbb{R} \to \mathbb{R}$. By Theorem 4.13, log is differentiable with

$$\log'(x) = \frac{1}{\exp'(\log x)} = \frac{1}{\exp(\log x)} = \frac{1}{x} \qquad \forall x \in (0, +\infty).$$

(2) Let $a \in \mathbb{C}$ and let $f : (0, +\infty) \to \mathbb{C}$, $f(x) = x^a$. Since by definition $f(x) = \exp(a \log x)$, it follows from the previous example and the chain rule that

$$f'(x) = \exp'(a \log x) \cdot \frac{a}{x} = \exp(a \log x) \cdot \frac{a}{x} = \frac{x^a}{x} = a x^{a-1}.$$

Hence, the formulas from Theorem 4.3.b) and Example 4.10.(1) generalize to arbitrary exponents.

(3) $\arctan : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ is the inverse function of $\tan : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$. By Theorem 4.13 and Example 4.10, arctan is differentiable with

$$\arctan'(x) = \frac{1}{\tan'(\arctan x)} = \cos^2(\arctan x) \qquad \forall x \in \mathbb{R}.$$

We want to simplify this expression. Fix $x \in \mathbb{R}$ and put $y := \arctan x$. Then

$$x^2 = (\tan y)^2 = \frac{\sin^2(y)}{\cos^2(y)} = \frac{1 - \cos^2(y)}{\cos^2(y)} = \frac{1}{\cos^2(y)} - 1.$$

Hence, $\cos^2(\arctan x) = \cos^2(y) = \frac{1}{1+x^2}$, which yields

$$\arctan'(x) = \frac{1}{1+x^2} \qquad \forall x \in \mathbb{R}.$$

(4) $\arcsin : [-1,1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the inverse function of $\sin|_{[-\frac{\pi}{2},\frac{\pi}{2}]}$. Since *sin* is differentiable with $\sin'(x) = \cos x \neq 0$ for all $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$, it follows from Theorem 4.13, arcsin is differentiable in each $x \in (-1,1)$ with

$$\arcsin'(x) = \frac{1}{\sin'(\arcsin(x))} = \frac{1}{\cos(\arcsin(x))} \qquad \forall x \in (-1,1).$$

Since $\cos(a)$ is positive for all $a \in (-\frac{\pi}{2}, \frac{\pi}{2})$, we derive that

$$\cos(\arcsin(x)) = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1-x^2}.$$

Hence,

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}} \qquad \forall x \in (-1,1).$$

Along the same lines one shows that $\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}$ for all $x \in (-1,1)$.

4.3 Local extrema and the mean value theorem

We want to begin this section by investigating the connection of derivatives to minima and maxima of functions. Same as continuity, one realizes that differentiability of a function at a point is a *local* notion, i.e. *the derivative of a function at a point only depends on the behaviour of the function in a neighborhood of the point* and not on the function as a whole.
As a consequence, it is too much to expect that derivatives can be used to study minima and maxima of functions on their whole domains. However, we can indeed use derivatives to detect so-called local extrema which we shall introduce next.

Definition 4.15. Let $D \subset \mathbb{C}$, $f : D \to \mathbb{R}$ and $x_0 \in D$. We say that *f has a local maximum (local minimum) in x_0* if there exists a D-neighborhood A of x_0, such that $f|_A$ has a maximum (minimum) in x_0. The value $f(x_0) \in \mathbb{R}$ is then called a local maximum (local minimum) of f and x_0 is called a local maximum point (local minimum point) of f. We further say that *f has a local extremum in x_0* if f has a local maximum in x_0 or a local minimum in x_0.

Obviously, every maximum (minimum) of a function is a local maximum (minimum).

Remark 4.16. By definition of a maximum (minimum) and of a D-neighborhood, one observes that f has a local maximum (local minimum) in x_0 if and only if there exists $\delta > 0$, such that

$$f(x) \leq f(x_0) \quad \forall x \in B_\delta(x_0) \cap D \qquad \Big(f(x) \geq f(x_0) \quad \forall x \in B_\delta(x_0) \cap D\Big).$$

Theorem 4.17 (Fermat)**.** *Let I be an open interval and let $f : I \to \mathbb{R}$ be differentiable. If f has a local extremum in $a \in I$, then $f'(a) = 0$.*

Proof. We assume w.l.o.g. that a is a local maximum point of f. (Otherwise, consider $-f$ instead of f.) Consider the continuous function $q : I \setminus \{a\} \to \mathbb{R}$,

$$q(x) := \frac{f(x) - f(a)}{x - a}.$$

Since I is open, there exists a $\delta > 0$ with $(a - \delta, a + \delta) \subset I$ and such that $f(x) \leq f(a)$ for all $x \in (a - \delta, a + \delta)$.
For $x \in (a, a + \delta)$, it follows that $q(x) \leq 0$. Likewise, for $x \in (a - \delta, a)$ we obtain $q(x) \geq 0$. Since by assumption, $\lim_{x \to a} q(x) = f'(a)$ exists, it follows from these inequalities that $f'(a) = 0$. □

Remark 4.18. The converse of Theorem 4.17 is false in general. For example, $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$, satisfies $f(0) = 0$, but 0 is neither a local minimum point nor a local maximum point of f, since $f|_{(-\delta,\delta)}$ has both positive and negative values for all $\delta > 0$.

As the previous example shows, Fermat's theorem provides a *necessary* condition on a point to be a local extremum point, but *not a sufficient* one. Later on in this chapter, we will see criteria that ensure that a point with $f'(x) = 0$ is indeed a local extremum point.
The following theorem presents a criterion for the existence of points with vanishing derivative.

Theorem 4.19 (Rolle[2])**.** *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ be differentiable. If $f(a) = f(b)$, then there exists $\xi \in (a, b)$ with $f'(\xi) = 0$.*

Proof. If f is constant, then the claim follows by Theorem 4.3.a). Assume that f is non-constant. Since $[a, b]$ is compact and f is continuous, f has a maximum and a minimum by Theorem 3.32, i.e. there exist $x_0, x_1 \in [a, b]$, such that

$$f(x_0) \leq f(x) \leq f(x_1) \qquad \forall x \in [a, b].$$

Since $f(a) = f(b)$, but $f(x_0) < f(x_1)$ by assumption, it follows that $x_0 \in (a, b)$ or $x_1 \in (a, b)$. In particular, there exists $\xi \in (a, b)$, such that f has a local extremum in ξ and it follows from Fermat's theorem that $f'(\xi) = 0$. □

Rolle's theorem is used in the proof of the following important theorem whose consequences and applications we will discuss throughout the rest of this section.

Theorem 4.20 (Mean value theorem)**.** *Let $a, b \in \mathbb{R}$ with $a < b$. Let $f : [a, b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ be differentiable. Then there exists $\xi \in (a, b)$ with*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Consider the function $g : [a, b] \to \mathbb{R}$,

$$g(x) = f(x) - (x - a)\frac{f(b) - f(a)}{b - a}.$$

Clearly, $g(a) = f(a)$, while

$$g(b) = f(b) - (b - a)\frac{f(b) - f(a)}{b - a} = f(b) - (f(b) - f(a)) = f(a).$$

[2]This theorem might look quite unspectacular and there is even a nice xkcd comic about it: `https://xkcd.com/2042/`. However, its use should not be underestimated.

Hence, $g(a) = g(b)$, so by Rolle's theorem, there exists $\xi \in (a, b)$ with

$$0 = g'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{b - a},$$

which shows the claim. □

Remark 4.21. The mean value theorem has an insightful geometric interpretation, connected to Remark 4.2. The ratio $\frac{f(b)-f(a)}{b-a}$ is the slope of the secant line of the graph of f that is going through $(a, f(a))$ and $(b, f(b))$ in $\mathbb{R}^2$. The mean value theorem says that there exists some point along the graph at which the tangent line to the graph has the same slope as this secant line. (See the illustration in the lecture videos.)

The mean value theorem as a lot of interesting and not at all obvious consequences. The first one connects the monotonicity of differentiable function with their derivatives.

Corollary 4.22. *Let $f : [a, b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ be differentiable. Then:*

a) f is monotonically increasing if and only if $f'(x) \geq 0$ for all $x \in (a, b)$.

b) f is monotonically decreasing if and only if $f'(x) \leq 0$ for all $x \in (a, b)$.

c) If $f'(x) > 0$ for all $x \in (a, b)$, then f is strictly monotonically increasing.

d) If $f'(x) < 0$ for all $x \in (a, b)$, then f is strictly monotonically decreasing.

Proof. a) "⇒": If f is monotonically increasing and $x_0 \in (a, b)$, then one checks that

$$q(x) := \frac{f(x) - f(x_0)}{x - x_0} \geq 0 \qquad \forall x \in [a, b] \setminus \{x_0\}.$$

Hence, $f'(x_0) = \lim_{x \to x_0} q(x) \geq 0$.

"⇐": Let $x_1, x_2 \in [a, b]$ with $x_1 < x_2$. By the mean value theorem, there exists $\xi \in (x_1, x_2)$ with

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(\xi) \geq 0. \tag{4.2}$$

Since $x_2 - x_1 > 0$, this implies that $f(x_2) - f(x_1) \geq 0$, i.e. $f(x_1) \leq f(x_2)$. This shows that f is monotonically increasing.

b) This is shown along the same lines as a).

c) If $f'(x) > 0$ for all $x \in (a, b)$, then the argument from (4.2) in part a) shows that $\frac{f(x_2)-f(x_1)}{x_2-x_1} > 0$ if $x_1 < x_2$, which then implies the strict monotonicity of f.

d) This is shown along the same lines as c).

□

Remark 4.23. The converses of parts c) and d) of Corollary 4.22 are false in general. The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$, is strictly monotonically increasing, but $f'(x) = 0$, while $-f$ is strictly monotonically decreasing with $-f'(0) = 0$.

It further follows from the mean value theorem that differentiability implies Lipschitz continuity under certain conditions.

Theorem 4.24. *Let $f : [a,b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ be differentiable. If there exists $L \geq 0$ with $|f'(x)| \leq L$ for all $x \in (a,b)$, then*

$$|f(x) - f(y)| \leq L \cdot |x - y| \qquad \forall x, y \in [a,b],$$

i.e. f is Lipschitz-continuous with Lipschitz constant L.

Proof. Let $x, y \in [a,b]$ with $x < y$. Then by the mean value theorem, there exists $\xi \in (x,y)$ with

$$\left| \frac{f(y) - f(x)}{y - x} \right| = |f'(\xi)| \leq L,$$

which implies $|f(x) - f(y)| \leq L \cdot |x - y|$ and $|f(y) - f(x)| \leq L \cdot |y - x|$. This proves the claim. □

This theorem has several interesting consequences about functions whose derivatives are related.

Corollary 4.25. *a) Let $f : [a,b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ be differentiable. Then f is constant if and only if $f'(x) = 0$ for all $x \in (a,b)$.*

b) Let $f, g : [a,b] \to \mathbb{R}$ be continuous and $f|_{(a,b)}$ and $g|_{(a,b)}$ be differentiable with

$$f'(x) = g'(x) \qquad \forall x \in (a,b).$$

Then there exists $c \in \mathbb{R}$ with $g(x) = f(x) + c$ for all $x \in (a,b)$.

c) Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable and assume that there exists $c \in \mathbb{R}$ with

$$f'(x) = cf(x) \qquad \forall x \in \mathbb{R}.$$

Then

$$f(x) = Ae^{cx} \qquad \forall x \in \mathbb{R},$$

where $A := f(0)$.

Proof. a) "⇒": This was shown in Theorem 4.3.a).

"⇐": If $|f'(x)| = 0$ for all $x \in (a,b)$, then it follows from Theorem 4.24 that $|f(x) - f(y)| \leq 0$ for all $x, y \in [a,b]$. Hence, $f(x) = f(y)$ for all $x, y \in [a,b]$, i.e. f is constant.

b) This follows by applying a) to $F : [a,b] \to \mathbb{R}$, $F(x) := g(x) - f(x)$.

c) Consider the function $F : \mathbb{R} \to \mathbb{R}$, $F(x) = f(x)e^{-cx}$. By Theorem 4.9.b), F is differentiable with

$$F'(x) = f'(x)e^{-cx} - f(x)(-ce^{-cx}) = (f'(x) - cf(x))e^{-cx} = 0 \qquad \forall x \in \mathbb{R}.$$

By part a), it follows that F is constant. Since $F(0) = f(0) = A$, it follows that $A = F(x) = f(x)e^{-cx}$ for all $x \in \mathbb{R}$. Multiplication by e^{cx} shows the claim.

□

Remark 4.26. Note that we can use Corollary 4.25.c) to give an alternative characterization of the real exponential function: $\exp : \mathbb{R} \to \mathbb{R}$ *is the unique differentiable function $f : \mathbb{R} \to \mathbb{R}$ that satisfies $f'(x) = f(x)$ for all $x \in \mathbb{R}$ and $f(0) = 1$.*

There is a more general version of the mean value theorem that we shall discuss next. Our main purpose of introducing it is its application to the subsequent proof of l'Hospital's rule that can be used to compute certain intricate limits of functions with the help of derivatives.

Theorem 4.27 (Generalized mean value theorem)**.** *Let $f, g : [a, b] \to \mathbb{R}$ be continuous and let $f|_{(a,b)}$ and $g|_{(a,b)}$ be differentiable. If $g'(x) \neq 0$ for all $x \in (a, b)$, then there exists $\xi \in (a, b)$ such that*

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Proof. Consider the function $h : [a, b] \to \mathbb{R}$,

$$h(x) = (g(b) - g(a))f(x) - (f(b) - f(a))g(x).$$

Then h is continuous, $h|_{(a,b)}$ is differentiable and

$$\begin{aligned} h(a) &= (g(b) - g(a))f(a) - (f(b) - f(a))g(a) = g(b)f(a) - f(b)g(a), \\ h(b) &= (g(b) - g(a))f(b) - (f(b) - f(a))g(b) = -g(a)f(b) + f(a)g(b), \end{aligned}$$

so $h(a) = h(b)$. By Rolle's theorem, there exists $\xi \in (a, b)$ with

$$0 = h'(\xi) = (g(b) - g(a))f'(\xi) - (f(b) - f(a))g'(\xi).$$

Since $g'(x) \neq 0$ for all $x \in (a, b)$, it follows from the mean value theorem that $g(a) \neq g(b)$, hence the above implies

$$(g(b) - g(a))f'(\xi) = (f(b) - f(a))g'(\xi) \quad \Leftrightarrow \quad \frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

□

Theorem 4.28 (L'Hospital's rule)**.** *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f, g : (a, b) \to \mathbb{R}$ be differentiable with $g'(x) \neq 0$ for all $x \in (a, b)$. Assume that one of the following holds:*

(i) $\lim_{x\to a} f(x) = 0$ *and* $\lim_{x\to a} g(x) = 0$,

(ii) $\lim_{x\to a} f(x) = +\infty$ *and* $\lim_{x\to a} g(x) = +\infty$,

If $\frac{f'}{g'}$ has a limit in a, then $\frac{f}{g}$ has a limit in a with

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

The analogous statement holds for limits in b, in $+\infty$ and in $-\infty$.

Proof. We first consider case (i) and consider f and g as defined on $[a, b)$ with $f(a) = 0$ and $g(a) = 0$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in (a, b) with $\lim_{n\to\infty} x_n = a$. By Theorem 4.27, for each $n \in \mathbb{N}$ there exists $\xi_n \in (a, x_n)$ with

$$\frac{f'(\xi_n)}{g'(\xi_n)} = \frac{f(x_n)}{g(x_n)} \qquad \forall n \in \mathbb{N}.$$

By the squeeze theorem, $(\xi_n)_{n\in\mathbb{N}}$ converges to a. Thus, if $\lim_{x\to a}\frac{f'(x)}{g'(x)}$ exists, then by Theorem 3.49

$$\lim_{n\to\infty}\frac{f(x_n)}{g(x_n)}=\lim_{n\to\infty}\frac{f'(\xi_n)}{g'(\xi_n)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}.$$

Since $(x_n)_{n\in\mathbb{N}}$ was chosen arbitrarily, the claim follows from Theorem 3.49.

We move on to case (ii). *This part of the proof is not discussed in the lecture videos.*

Put $L:=\lim_{x\to a}\frac{f'(x)}{g'(x)}$ and let $\varepsilon>0$. Then there exists $\delta_1>0$ with

$$\left|\frac{f'(x)}{g'(x)}-L\right|<\frac{\varepsilon}{2}\quad\forall x\in(a,a+\delta_1).$$

Let $x,y\in(a,a+\delta_1)$. Then by the generalized mean value theorem there exists $\xi\in(a,a+\delta_1)$ with $\frac{f'(\xi)}{g'(\xi)}=\frac{f(x)-f(y)}{g(x)-g(y)}$ and thus

$$\left|\frac{f(x)-f(y)}{g(x)-g(y)}-L\right|=\left|\frac{f'(\xi)}{g'(\xi)}-L\right|<\frac{\varepsilon}{2}\quad\forall x,y\in(a,a+\delta_1)\text{ with }x\neq y.$$

We further compute:

$$\frac{f(x)}{g(x)}=\frac{f(x)-f(y)}{g(x)-g(y)}\cdot\frac{f(x)(g(x)-g(y))}{g(x)(f(x)-f(y))}=\frac{f(x)-f(y)}{g(x)-g(y)}\cdot\frac{\frac{g(x)-g(y)}{g(x)}}{\frac{f(x)-f(y)}{f(x)}}=\frac{f(x)-f(y)}{g(x)-g(y)}\cdot\frac{1-\frac{g(y)}{g(x)}}{1-\frac{f(y)}{f(x)}}.$$

With $y\in(a,a+\delta_1)$ fixed, it holds by assumption that

$$\lim_{x\to a}\frac{1-\frac{g(y)}{g(x)}}{1-\frac{f(y)}{f(x)}}=1,$$

hence there exists $\delta_2>0$ with $\delta_2\leq y-a$, such that

$$\left|\frac{f(x)}{g(x)}-\frac{f(x)-f(y)}{g(x)-g(y)}\right|<\frac{\varepsilon}{2}\qquad\forall x\in(a,a+\delta_2)$$

Thus, if we put $\delta:=\min\{\delta_1,\delta_2\}$, it holds for all $x\in(a,a+\delta)$ that

$$\left|\frac{f(x)}{g(x)}-L\right|\leq\left|\frac{f(x)}{g(x)}-\frac{f(x)-f(y)}{g(x)-g(y)}\right|+\left|\frac{f(x)-f(y)}{g(x)-g(y)}-L\right|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

Since $\varepsilon>0$ was arbitrary, it follows that $\lim_{x\to a}\frac{f(x)}{g(x)}=L$, which shows the claim.
The case of $x\to b$ is proven completely analogously, while the cases of $x\to+\infty$ and $x\to-\infty$ follow by studying the limit of $\frac{f(\frac{1}{x})}{g(\frac{1}{x})}$ as $x\searrow 0$ and $x\nearrow 0$. We omit the details. □

Example 4.29. We want to compute $\lim\limits_{x\searrow 0}\left(\frac{1}{\sin x}-\frac{1}{x}\right)$. For all $x\in(0,\frac{\pi}{2})$,

$$\frac{1}{\sin x}-\frac{1}{x}=\frac{x-\sin x}{x\cdot\sin x}.$$

$x - \sin x$ and $g(x) = x \cdot \sin x$ are differentiable with

$$\lim_{x \searrow 0}(x - \sin x) = 0, \qquad \lim_{x \searrow 0} x \cdot \sin x = 0.$$

Moreover, by the product rule $g'(x) = \sin x + x \cos x \neq 0$ for all $x \in (0, \frac{\pi}{2})$. Thus, it follows by l'Hospital's rule that

$$\lim_{x \searrow 0}\left(\frac{1}{\sin x} - \frac{1}{x}\right) = \lim_{x \searrow 0} \frac{1 - \cos x}{\sin x + x \cos x}$$

assuming the limit on the right-hand side exists. We compute that $\lim_{x \searrow 0}(1 - \cos x) = 0$ and $\lim_{x \searrow 0}(\sin x + x \cos x) = 0$. Both $1 - \cos x$ and $h(x) := \sin x + x \cos x$ are differentiable with

$$h'(x) = \cos x + \cos x + x(-\sin x) = 2\cos x - x \sin x. \qquad \forall x \in \left(0, \frac{\pi}{2}\right)$$

Since $\lim_{x \searrow 0} h'(x) = 2\cos 0 - 0 = 2 \neq 0$, there exists $\frac{\pi}{2} > \delta > 0$ with $h'(x) \neq 0$ for all $x \in (0, \delta)$. Thus, it follows by another application of l'Hospital's rule that

$$\lim_{x \searrow 0} \frac{1 - \cos x}{\sin x + x \cos x} = \lim_{x \searrow 0} \frac{\sin x}{2\cos x - x \sin x} = \frac{\sin 0}{2\cos 0 - 0} = \frac{0}{2} = 0.$$

Thus, we have shown that

$$\lim_{x \searrow 0}\left(\frac{1}{\sin x} - \frac{1}{x}\right) = 0.$$

4.4 Higher order derivatives and Taylor's theorem

Next we want to iterate the differentiation procedure. Given a differentiable function $f : I \to \mathbb{C}$, we want to view the derivative as a function $f' : I \to \mathbb{C}$ and ask if this function is again differentiable. This is the starting point for considering derivatives of higher order which we shall define recursively.

Definition 4.30. Let I be an interval and and $f : I \to \mathbb{C}$ be differentiable. Put $f^{(0)} := f$.

a) We call $f' : I \to \mathbb{C}$ *the derivative*[3] *of* f. We also denote $f^{(1)} := f'$.

b) f is *twice differentiable in* $a \in I$ if $f' : I \to \mathbb{C}$ is differentiable in a. We then put $f''(a) := (f')'(a)$ and call $f''(a)$ *the second derivative of* f *in* a. We say that f is *twice differentiable* if it is twice differentiable in every $a \in I$. We also denote $f^{(2)} := f''$.

c) Let $k \in \mathbb{N}$ with $k \geq 3$. f is *k times differentiable in $a \in I$* if f is $k-1$ times differentiable and if $f^{(k-1)} : I \to \mathbb{C}$ is differentiable in a. In this case we put $f^{(k)}(a) := (f^{(k-1)})'(a)$ and call $f^{(k)}(a)$ the *k-th derivative* or *derivative of k-th order in a*. We say that f is *k times differentiable* if it is k times differentiable in every $a \in I$.

d) f is *smooth* if it is k times differentiable for all $k \in \mathbb{N}$.

We will occasionally denote the third and fourth derivative of a function f by f''' and f'''', resp.

[3] In case you are thinking "Didn't we already define this before?", the answer is yes and no. Yes, we have defined the derivative of f *in a point*, but we haven't formally defined "the" derivative as a function $I \to \mathbb{C}$.

Example 4.31. (1) Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^n$, where $n \in \mathbb{N}$. Then f is smooth and its k-th order derivative is given by

$$f^{(k)}(x) = n(n-1)\dots(n-k+1)x^{n-k} \qquad \text{if } k \leq n$$

and $f^{(k)}(x) = 0$ for all $k > n$.

(2) Consider $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \sin x$. Then $g'(x) = \cos x$ is again differentiable with $g''(x) = \cos'(x) = -\sin x$. One proves by induction that g is smooth with

$$f^{(2k)}(x) = (-1)^k \sin x, \qquad f^{(2k+1)}(x) = (-1)^k \cos x \qquad \forall k \in \mathbb{N}.$$

Even without new results, we can already derive an important inequality using higher derivatives.

Theorem 4.32 (Inequality of arithmetic and geometric means)**.** *For all $a_1, a_2, \dots, a_n \in [0, +\infty)$, where $n \in \mathbb{N}$, it holds that*

$$\sqrt[n]{a_1 a_2 \dots a_n} \leq \frac{a_1 + a_2 + \dots + a_n}{n}.$$

Moreover, equality holds in this inequality if and only if $a_1 = a_2 = \dots = a_n$. ($\sqrt[n]{a_1 a_2 \dots a_n}$ is called the geometric mean *of $a_1, \dots, a_n$, while $\frac{a_1+a_2+\dots+a_n}{n}$ is called the* arithmetic mean *of $a_1, \dots, a_n$.)*

To use a more concise name, we will refer to the inequality from Theorem 4.32 as the *AM-GM inequality*.

Remark 4.33. (1) In the case of $n = 2$, the AM-GM inequality can be proven much more elementarily. Let $a, b \in [0, +\infty)$. Then $0 \leq (a-b)^2 = a^2 - 2ab + b^2$. Adding $4ab$ to both sides of the inequality yields

$$4ab \leq a^2 + 2ab + b^2 \quad \Leftrightarrow \quad ab \leq \frac{(a+b)^2}{4}.$$

Taking roots shows that

$$\sqrt{ab} \leq \frac{a+b}{2}.$$

(2) The arithmetic mean of n numbers is the number we usually consider as the "average" of these n numbers, e.g. in computing the average grade of an exam or the average price of an item that is available to buy in n different shops.

The geometric mean is less intuitive, but usually occurs in considering growth rates. For example, consider a population of a species that is observed over n years. Assume that the population grows or shrinks by a factor of a_i in year i. Then after n years, the average growth rate a of the population would be the geometric mean of $a_1, \dots, a_n$ in the sense that the population after n years would be the same if the growth factor would have been constantly equal to a in each year.

Proof of Theorem 4.32. We will prove the claim by induction over n.
The base case of $n = 1$ is obviously nothing more than $a_1 \leq a_1$ for all $a_1 \in [0, +\infty)$, which is obviously true.
For the induction step, we assume that for a certain $n \in \mathbb{N}$ the inequality

$$\sqrt[n]{a_1 a_2 \dots a_n} \leq \frac{a_1 + a_2 + \dots + a_n}{n}$$

holds for all $a_1, \dots, a_n \in [0, +\infty)$ with equality if and only if $a_1 = a_2 = \cdots = a_n$. We want to derive that

$$\sqrt[n+1]{a_1 a_2 \dots a_{n+1}} \overset{!}{\leq} \frac{a_1 + a_2 + \cdots + a_{n+1}}{n+1} \quad \forall a_1, \dots, a_{n+1} \in [0, +\infty)$$

with equality if and only if $a_1 = a_2 = \cdots = a_{n+1}$.
If $a_i = 0$ for some $i \in \{1, 2, \dots, n+1\}$, then the claim is obvious, so it suffices to consider $a_1, \dots, a_n, a_{n+1} \in (0, +\infty)$. We choose and fix positive numbers $a_1, \dots, a_n$ and consider the function

$$\begin{aligned} f : (0, +\infty) \to \mathbb{R}, \qquad f(x) &= \frac{a_1 + \cdots + a_n + x}{n+1} - \sqrt[n+1]{a_1 \dots a_n x} \\ &= \frac{a_1 + \cdots + a_n + x}{n+1} + (a_1 \dots a_n)^{\frac{1}{n+1}} x^{\frac{1}{n+1}}. \end{aligned}$$

We want to show that $f(x) \geq 0$ for all $x \in (0, +\infty)$. We observe using our previous computations and the chain rule that f is twice differentiable with

$$f'(x) = \frac{1}{n+1} - \frac{1}{n+1} (a_1 \dots a_n)^{\frac{1}{n+1}} x^{-\frac{n}{n+1}}$$

and

$$f''(x) = \frac{n}{(n+1)^2} (a_1 \dots a_n)^{\frac{1}{n+1}} x^{-\frac{2n+1}{n+1}},$$

which particularly shows that $f''(x) > 0$ for all $x \in (0, +\infty)$. Thus, by Corollary 4.22.c), $f' : (0, +\infty) \to \mathbb{R}$ is strictly monotonically increasing. In particular, there is at most one point $x_0 \in (0, +\infty)$ with $f'(x_0) = 0$. More precisely, we compute

$$\begin{aligned} 0 = f'(x_0) &= \frac{1}{n+1} - \frac{1}{n+1} (a_1 \dots a_n)^{\frac{1}{n+1}} x_0^{-\frac{n}{n+1}} \\ \Leftrightarrow \quad 0 &= 1 - (a_1 \dots a_n)^{\frac{1}{n+1}} x_0^{-\frac{n}{n+1}} \\ \Leftrightarrow \quad x_0^{\frac{n}{n+1}} &= (a_1 \dots a_n)^{\frac{1}{n+1}} \\ \Leftrightarrow \quad x_0 &= (a_1 \dots a_n)^{\frac{1}{n}} = \sqrt[n]{a_1 \dots a_n}. \end{aligned}$$

By the strict monotonicity of f', it follows that

$$\begin{aligned} f'(x) &< 0 \qquad \text{if } x \in (0, x_0), \\ f'(x) &> 0 \qquad \text{if } x \in (x_0, +\infty). \end{aligned}$$

By Corollary 4.22.c) and d), this yields that $f|_{(0,x_0]}$ is strictly monotonically decreasing while $f|_{[x_0,+\infty)}$ is strictly monotonically increasing. Both of these observations combined show that f

attains its minimum in x_0 with $f(x) > f(x_0)$ if $x \neq x_0$. We compute that

$$\begin{aligned} f(x_0) &= \frac{a_1 + \cdots + a_n + (a_1 \dots a_n)^{\frac{1}{n}}}{n+1} - \left(a_1 \dots a_n (a_1 \dots a_n)^{\frac{1}{n}}\right)^{\frac{1}{n+1}} \\ &= \frac{a_1 + \cdots + a_n + (a_1 \dots a_n)^{\frac{1}{n}}}{n+1} - (a_1 \dots a_n)^{\frac{1}{n+1}} (a_1 \dots a_n)^{\frac{1}{n(n+1)}} \\ &= \frac{a_1 + \cdots + a_n + (a_1 \dots a_n)^{\frac{1}{n}}}{n+1} - (a_1 \dots a_n)^{\frac{n+1}{n(n+1)}} \\ &= \frac{a_1 + \cdots + a_n}{n+1} + \left(\frac{1}{n+1} - 1\right) (a_1 \dots a_n)^{\frac{1}{n}} \\ &= \frac{n}{n+1} \left(\frac{a_1 + \cdots + a_n}{n} - (a_1 \dots a_n)^{\frac{1}{n}}\right). \end{aligned}$$

Hence, by the induction hypothesis, $f(x_0) \geq 0$ and thus $f(x) \geq 0$ for all $x \in (0, +\infty)$. Moreover, by the induction hypothesis $f(x_0) = 0$ if and only if $a_1 = a_2 = \cdots = a_n$ in which case it follows that $x_0 = a_1$ as well. This completes the induction step, hence the proof of the claim. □

As we have carried out in the previous sections, the purpose of the derivative of a function to approximate a function near a particular point by an affine-linear function, i.e. a polynomial of degree one. We want to generalize this idea.
In the same spirit as before, we next want to use higher order derivatives *to locally approximate an n times differentiable function near a point by a polynomial of degree n*.

As a guideline and as a motivational idea we assume that $f : I \to \mathbb{C}$ is n times differentiable, take $a \in I$ and want to find a polynomial $P_n : \mathbb{R} \to \mathbb{C}$, $P_n(x) = \sum_{r=0}^{n} c_r (x-a)^r$, where $c_0, \dots, c_n \in \mathbb{C}$, such that

$$f^{(k)}(a) = P_n^{(k)}(a) \qquad \forall k \in \{0, 1, \dots, n\}.$$

One derives from Example 4.31.(1) and the usual differentiation rules that

$$P_n^{(k)}(x) = \sum_{r=k}^{n} c_r r(r-1)\dots(r-k+1)(x-a)^{r-k},$$

for all $k \in \{0, 1, \dots, n\}$, which particularly implies that

$$P_n^{(k)}(a) = c_k k!.$$

Hence, to obtain $P_n^{(k)}(a) = f^{(k)}(a)$ for all $k \in \{0, 1, \dots, n\}$, we need to put $c_k = \frac{f^{(k)}(a)}{k!}$ for each k.

Definition 4.34. Let I be an interval, let $f : I \to \mathbb{C}$ be an n times differentiable function and let $a \in I$. The *n-th Taylor polynomial of f in a* is given by

$$T_n f(x, a) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k.$$

Example 4.35. Let $f : (-\infty, 0) \to \mathbb{R}$, $f(x) = \frac{1}{x^2} = x^{-2}$.

$$f'(x) = -2x^{-3}, \quad f''(x) = (-3)(-2)x^{-4}, \quad \dots,$$

and one proves by induction that f is smooth with

$$f^{(k)}(x) = (-1)^k(k+1)! \cdot x^{-k-2} = (-1)^k \frac{(k+1)!}{x^{k+2}}. \qquad \forall x \in (-\infty, 0),\ k \in \mathbb{N}.$$

Let $a = -1$. Then $f(-1) = 1$ and

$$f^{(k)}(-1) = (-1)^k \frac{(k+1)!}{(-1)^{k+2}} = (k+1)! \qquad \forall k \in \mathbb{N}.$$

Thus,

$$\begin{aligned}
T_1 f(x, -1) &= 1 + \frac{2!}{1!}(x+1) = 1 + 2(x+1) = 2x + 3, \\
T_2 f(x, -1) &= 1 + 2(x+1) + \frac{3!}{2!}(x+1)^2 = 1 + 2(x+1) + 3(x+1)^2 = 3x^2 + 8x + 6, \\
T_n f(x, -1) &= \sum_{k=0}^{n} (k+1)(x+1)^k \qquad \forall n \in \mathbb{N}.
\end{aligned}$$

The following fundamental theorem shows that the Taylor polynomials of a function in a point are the best approximations of the function by polynomials in that point.

Theorem 4.36 (Taylor)**.** *Let I be an interval, $a \in I$ and let $f : I \to \mathbb{C}$ be n times differentiable in a, where $n \in \mathbb{N}$. Then*

$$f(x) = T_n f(x, a) + \varphi_n(x),$$

where $\varphi_n : I \to \mathbb{C}$ satisfies

$$\lim_{x \to a} \frac{\varphi_n(x)}{(x-a)^n} = 0.$$

Before we prove this theorem, we prove a helpful statement used in its proof.

Lemma 4.37. [4] *Let I be an interval, $a \in I$ and let $f : I \to \mathbb{C}$ be n times differentiable in a, where $n \in \mathbb{N}$, with $f^{(k)}(a) = 0$ for all $k \in \{0, 1, \ldots, n\}$. Then*

$$\lim_{x \to a} \frac{f^{(n-k)}(x)}{(x-a)^k} = 0 \qquad \forall k \in \{1, 2, \ldots, n\}$$

Proof. We will show the claim by induction over k.

Base case: Since $f^{(n-1)}$ is differentiable in a, it follows from Theorem 4.6 that

$$f^{(n-1)}(x) = f^{(n-1)}(a) + f^{(n)}(a)(x-a) + r(x) \qquad \forall x \in I,$$

where $r : I \to \mathbb{C}$ satisfies $\lim_{x \to a} \frac{r(x)}{x-a} = 0$. Since $f^{(n-1)}(a) = f^{(n)}(a) = 0$, it follows that $f^{(n-1)}(x) = r(x)$, which immediately proves the case of $k = 1$.

Induction step: Assume that for some $k < n$ it holds that

$$\lim_{x \to a} \frac{f^{(n-k)}(x)}{(x-a)^k} = 0. \tag{4.3}$$

[4] A *lemma* is a mathematical statement that is not all that interesting in its own right, but that is mostly used to prove another more important statement afterwards.

Let $(x_n)_{n\in\mathbb{N}}$ be an arbitrary sequence in I with $\lim_{n\to\infty} x_n = a$. By the mean value theorem, for each $n \in \mathbb{N}$ there exists $\xi_n \in (a, x_n) \cup (x_n, a)$ with

$$f^{(n-k-1)}(x_n) = f^{(n-k-1)}(x_n) - f^{(n-k-1)}(a) = f^{(n-k)}(\xi_n)(x_n - a).$$

One checks without difficulties that $\lim_{n\to\infty} \xi_n = a$ as well. Thus,

$$\lim_{n\to\infty} \left| \frac{f^{(n-k-1)}(x_n)}{(x_n-a)^{k+1}} \right| = \lim_{n\to\infty} \frac{|f^{(n-k)}(\xi_n)|}{|x_n-a|^k} \leq \lim_{n\to\infty} \frac{|f^{(n-k)}(\xi_n)|}{|\xi_n-a|^k} = \left| \lim_{n\to\infty} \frac{f^{(n-k)}(\xi_n)}{(\xi_n-a)^k} \right| \overset{(4.3)}{=} 0.$$

Since $(x_n)_{n\in\mathbb{N}}$ was arbitrary, it follows that $\lim_{x\to a} \frac{f^{(n-k-1)}(x)}{(x-a)^{k+1}} = 0$, which completes the induction step. □

Proof of Theorem 4.36. Define $P_n(x) := T_n f(x, a)$ and $\varphi_n(x) := f(x) - P_n(x)$ for all $x \in I$. Then φ_n is n times differentiable in a and by the linearity of derivatives,

$$\varphi_n^{(k)}(a) = f^{(k)}(a) - P_n^{(k)}(a) \qquad \forall x \in I,\ k \in \{1, 2, \dots, n\}.$$

By the above computations, $f^{(k)}(a) = P_n^{(k)}(a)$ for all $k \in \{1, 2, \dots, n\}$, hence

$$\varphi_n^{(k)}(a) = 0 \qquad \forall k \in \{1, 2, \dots, n\}.$$

The claim follows immediately from the case $k = n$ of Lemma 4.37. □

Remark 4.38. (1) The version of Taylor's theorem that is Theorem 4.36 is also called the *Péano form* of Taylor's theorem.

(2) The limit property of φ_n in Taylor's theorem is the key of the statement: it tells us that the n-th Taylor polynomial of f in a is the *best local approximation of f near a by a polynomial of degree n*, since everything that remains goes faster to 0 than $(x-a)^n$ as $x \to a$.

(3) Unfortunately, even the best approximation of f by a polynomial does not need to be a *good* approximation as the following example shows: Let

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} e^{-\frac{1}{x^2}} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

and let $a = 0$. One can check (exercise) that f is indeed smooth and satisfies $f^{(k)}(0) = 0$ for all $k \in \mathbb{N}$. In particular, $T_n f(x, 0) = 0$ for all $n \in \mathbb{N}$, $x \in \mathbb{R}$, which shows that the Taylor polynomials of f at 0 don't tell us anything about the behaviour of f near 0.

4.5 Computations and examples of Taylor's theorem

In the following example we study two important examples of how to approximate complicated functions by polynomials.

Example 4.39. (1) Consider $f : (-1, +\infty) \to \mathbb{R}$, $f(x) = \log(1 + x)$. Since $f'(x) = \frac{1}{1+x}$ is a rational function, i.e. a quotient of polynomials, one checks that f is indeed smooth. Up to third order, its derivatives are computed by the quotient rule as

$$f'(x) = \frac{1}{1+x}, \quad f''(x) = -\frac{1}{(1+x)^2}, \quad f'''(x) = \frac{2}{(1+x)^3}.$$

One shows by induction that

$$f^{(k)}(x) = (-1)^{k-1}\frac{(k-1)!}{(1+x)^k} \quad \forall k \in \mathbb{N}, x \in (-1, +\infty).$$

In particular,

$$f(0) = 0, \quad f^{(k)}(0) = (-1)^{k-1}(k-1)! \quad \forall k \in \mathbb{N},$$

so the Taylor polynomials of f in 0 have the form

$$T_n f(x, 0) = \sum_{k=1}^{n} \frac{(-1)^{k-1}(k-1)!}{k!}x^k = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n}.$$

Hence, by Taylor's theorem,

$$\log(1+x) = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k + r_n(x) \qquad \forall x \in (-1, +\infty),$$

where $r_n : (-1, +\infty) \to \mathbb{R}$ satisfies $\lim_{x\to 0} \frac{r_n(x)}{x^n} = 0$.

(2) We want to study the Taylor polynomials of $\arctan : \mathbb{R} \to \mathbb{R}$ in 0. By Example 4.14.(3), arctan is differentiable with

$$\arctan'(x) = \frac{1}{1+x^2} \qquad \forall x \in \mathbb{R}.$$

Similar to part (1), one concludes that arctan is indeed smooth. To obtain a general formula for its k-th derivative, we need to employ a little trick We interpret $\arctan'$ as a complex function by computing

$$\arctan'(x) = \frac{1}{(1+ix)(1-ix)} = \frac{1}{2}\Big(\frac{1-ix\ +\ 1+ix}{(1+ix)(1-ix)}\Big) = \frac{1}{2}\Big(\frac{1}{1+ix} + \frac{1}{1-ix}\Big).$$

From the usual rules, we compute that

$$\arctan''(x) = \frac{1}{2}\left(\frac{-i}{(1+ix)^2} + \frac{i}{(1-ix)^2}\right),$$
$$\arctan'''(x) = \frac{1}{2}\left(\frac{(-2i)(-i)}{(1+ix)^3} + \frac{2i\cdot i}{(1-ix)^3}\right).$$

More generally, one proves by induction that

$$\arctan^{(k)}(x) = \frac{(k-1)!}{2}\left(\frac{(-i)^{k-1}}{(1+ix)^k} + \frac{i^{k-1}}{(1-ix)^k}\right) \qquad \forall k \in \mathbb{N},\ x \in \mathbb{R}.$$

In particular,

$$\arctan^{(k)}(0) = \frac{(k-1)!}{2}((-i)^{k-1} + i^{k-1}) = \frac{(k-1)!\cdot i^{k-1}}{2}((-1)^{k-1} + 1).$$

This shows that $\arctan^{(2n)}(0) = 0$ for all $n \in \mathbb{N}$, while

$$\arctan^{(2n+1)}(0) = \frac{(2n)!\cdot i^{2n}}{2}((-1)^{2n} + 1) = (2n)!(-1)^n \qquad \forall n \in \mathbb{N}_0.$$

Thus, for each $n \in \mathbb{N}_0$ we obtain the following Taylor polynomials of $f = \arctan$ in 0:

$$T_{2n+1}f(x,0) = \sum_{k=0}^{n} \frac{(2k)!(-1)^k}{(2k+1)!}x^{2k+1} = \sum_{k=0}^{n} \frac{(-1)^k}{2k+1}x^{2k+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1}.$$

Consequently, for each $n \in \mathbb{N}_0$,

$$\arctan x = \sum_{k=0}^{n} \frac{(-1)^k}{2k+1}x^{2k+1} + r_{2n+1}(x),$$

where $r_{2n+1} : \mathbb{R} \to \mathbb{R}$ satisfies $\lim_{x\to 0} \frac{r_{2n+1}(x)}{x^{2n+1}} = 0$.

Remark 4.40. Further examples can be obtained from power series. As we shall see in the next semester, if $f : (-r, r) \to \mathbb{C}$, $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is well-defined, then f is differentiable and

$$T_n f(x,0) = \sum_{k=0}^{n} a_k x^k \qquad \forall n \in \mathbb{N}.$$

Taylor's theorem can be used to compute limits as the following example shows.

Example 4.41. We want to compute

$$\lim_{x\to 0} \frac{\arctan x - x}{x^2} \quad \text{and} \quad \lim_{x\to 0} \frac{\arctan x - x}{x^3}.$$

As seen in Example 4.39,

$$\arctan x = x - \frac{x^3}{3} + r_3(x),$$

where $\lim_{x\to 0} \frac{r_3(x)}{x^3} = 0$. Thus,

$$\lim_{x\to 0} \frac{\arctan x - x}{x^3} = \lim_{x\to 0} \frac{-\frac{x^3}{3} + r_3(x)}{x^3} = \lim_{x\to 0} \left(-\frac{1}{3} + \frac{r_3(x)}{x^3}\right) = -\frac{1}{3}$$

and

$$\lim_{x\to 0} \frac{\arctan x - x}{x^2} = \lim_{x\to 0} \frac{-\frac{x^3}{3} + r_3(x)}{x^2} = \lim_{x\to 0} \left(-\frac{1}{3}x + x\frac{r_3(x)}{x^3}\right) = 0.$$

Note that we still don't know if the Taylor polynomials from Example 4.39 are good approximations of $\log(1+x)$ and $\arctan x$, resp. To answer questions like these, we need to estimate the difference between a function and its Taylor polynomials. For this purpose, we establish the following version of Taylor's theorem.

Theorem 4.42 (Lagrange remainder form)**.** *Let I be an interval, $a \in I$, $n \in \mathbb{N}_0$ and let $f : I \to \mathbb{C}$ be $n+1$ times differentiable. Then for each $x \in I$ there exists $\xi \in I$ between a and x with*

$$f(x) = T_n f(x,a) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}.$$

Proof. For $x = a$ the statement is obvious, so we assume $x \neq a$. We further assume w.l.o.g. that $x > a$, the other case is proven completely analogously. Let $M \in \mathbb{R}$ be uniquely defined by

$$f(x) = T_n f(x,a) + M(x-a)^{n+1}.$$

Put $p_n(t) := T_n f(t,a)$ and consider

$$g : I \to \mathbb{C}, \qquad g(t) := f(t) - p_n(t) - M(t-a)^{n+1}.$$

Then g is $n+1$ times differentiable. Since p_n is a polynomial of degree n, the usual rules of differentiation show that

$$g^{(n+1)}(t) = f^{(n+1)}(t) - M(n+1)!(t-a)^0 = f^{(n+1)}(t) - M(n+1)!.$$

Moreover,

$$g^{(k)}(a) = f^{(k)}(a) - p_n^{(k)}(a) = 0 \qquad \forall k \in \{0,1,\dots,n\}.$$

By choice of M, it further holds that $g(x) = 0$. Since $g(a) = 0$, it holds by Rolle's theorem that there exists $\xi_1 \in (a,x)$ with $g'(\xi_1) = 0$. Since $g'(a) = 0$, it again holds by Rolle's theorem that there exists $\xi_2 \in (a,\xi_1)$ with $g''(\xi_2) = 0$. Iterating this argument $n+1$ times, we obtain that there exists $\xi \in (a,x)$ with

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - M(n+1)! \qquad \Leftrightarrow \qquad M = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

which shows the claim. □

Corollary 4.43. *Let I be an interval, $n \in \mathbb{N}$ and let $f : I \to \mathbb{C}$ be $n+1$ times differentiable. If $f^{(n+1)}(x) = 0$ for all $x \in I$, then f is a polynomial of degree at most n.*

Proof. By Theorem 4.42, it holds under the asumption that $f(x) = T_n f(x,a)$, where $a \in I$ is chosen arbitrarily. □

We apply the Lagrange remainder formula to the functions studied in Example 4.39.

Theorem 4.44. *a) For all $x \in [0,1]$, it holds that*

$$\log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} x^n,$$

in particular

$$\log 2 = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \dots.$$

b) For all $x \in [-1,1]$ it holds that

$$\arctan x = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}.$$

In particular,

$$\pi = 4 \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots\right).$$

Proof. a) Let $f : [0,1] \to \mathbb{R}$, $f(x) = \log(1+x)$. Put $R_n(x) := f(x) - T_n f(x,0)$ for each $x \in [0,1]$. We want to show that $(R_n(x))_{n\in\mathbb{N}}$ converges to 0 for all $x \in [0,1]$. By Theorem 4.42 and the computations of Example 4.39.(1), for each $n \in \mathbb{N}$ and $x \in [0,1]$ there exists $\xi \in [0,1]$ with

$$|R_n(x)| = \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} x^{n+1} \right| = \frac{1}{|(n+1)(1+\xi)^{n+1}|} |x|^{n+1} \le \frac{1}{(n+1)(1+\xi)^{n+1}} \le \frac{1}{n+1}.$$

Hence, by the squeeze theorem, $\lim_{n\to\infty} |R_n(x)| = 0$ and thus $\lim_{n\to\infty} R_n(x) = 0$ for each $x \in [0,1]$, which shows the claim.

b) Let $f(x) = \arctan x$ and let $R_{2n+1}(x) := f(x) - T_{2n+1}f(x,0)$ for all $n \in \mathbb{N}$. By Theorem 4.42 and the computations of Example 4.39.(2), for each $n \in \mathbb{N}$ and $x \in [-1,1]$ there exists $\xi \in [-1,1]$ with

$$\begin{aligned}|R_{2n+1}(x)| &= \frac{|f^{(2n+2)}(\xi)|}{(2n+2)!}|x|^{2n+2} \leq \frac{1}{2(2n+2)}\left|\frac{(-i)^{2n+1}}{(1+i\xi)^{2n+2}} + \frac{i^{2n+1}}{(1-i\xi)^{2n+2}}\right| \\ &\leq \frac{1}{2(2n+2)}\left(\frac{|-i|^{2n+1}}{|1+i\xi|^{2n+2}} + \frac{|i|^{2n+1}}{|1-ix|^{2n+2}}\right) \\ &\leq \frac{1}{2(2n+2)}\left(\frac{1}{|1+i\xi|^{2n+2}} + \frac{1}{|1-ix|^{2n+2}}\right).\end{aligned}$$

Since $|1+i\xi| = |1-i\xi| = \sqrt{1+\xi^2} \geq 1$, it follows that $|R_{2n+1}(x)| \leq \frac{1}{2n+2}$ for all $x \in [-1,1]$ and $n \in \mathbb{N}$. This shows the claim in analogy with part a). To obtain the claimed identity for π, we observe that $\sin(\frac{\pi}{4}) = \cos(\frac{\pi}{4})$ and thus $\tan(\frac{\pi}{4}) = 1$ by Exercise 3 from Sheet 8. Hence, $\pi = 4\arctan 1$ and the identity follows from the proven formula for $\arctan x$ with $x = 1$.

□

Remark 4.45. The equation from of Theorem 4.44.a) actually holds true for all $x \in (-1,1]$, which requires some additional effort.

To conclude this section, we want to revisit local extrema of differentiable function. Fermat's theorem established a *necessary* condition for a point to be a local extremum point. With the help of Taylor's theorem we can now establish *sufficient* conditions for a point to be a local maximum point or a local minimum point.

Theorem 4.46. *Let I be an open interval, $a \in I$ and $n \in \mathbb{N}$ with $n \geq 2$. Let $f : I \to \mathbb{C}$ be n times differentiable. Assume that $f^{(n)} : I \to \mathbb{C}$ is continuous and that*

$$f'(a) = f''(a) = \cdots = f^{(n-1)}(a) = 0, \quad f^{(n)}(a) \neq 0.$$

a) If n is odd, then f does not have a local extremum in a.

b) If n is even and $f^{(n)}(a) < 0$, then f has a local maximum in a.

c) If n is even and $f^{(n)}(a) > 0$, then f has a local minimum in a.

Proof. Since $f^{(n)}(a) \neq 0$ and $f^{(n)}$ is continuous, it follows as in Exercise 1.a) of Sheet 6 that there exists $\delta > 0$, such that $f^{(n)}(x) \neq 0$ for all $x \in (a-\delta, a+\delta) \cap I$. Since I is open, we may choose δ so small that $(a-\delta, a+\delta) \subset I$. It follows directly from the assumptions that $T_{n-1}f(x,a) = f(a)$ for all $x \in I$, so by Theorem 4.42, for each $x \in (a-\delta, a+\delta)$ there exists $\xi \in (a-\delta, a+\delta)$ with

$$f(x) = f(a) + \frac{f^{(n)}(\xi)}{n!}(x-a)^n. \qquad \Leftrightarrow \qquad f(x) - f(a) = \frac{f^{(n)}(\xi)}{n!}(x-a)^n. \tag{4.4}$$

a) If n is odd, then $(x-a)^n$ is positive for $x > a$ and negative for $x < a$. Hence, by (4.4) the expression $f(x) - f(a)$ takes both positive and negative values in any I-neighborhood of a. Hence, a is neither a local maximum point nor a local minimum point of f.

b) By construction of δ, it follows that $f^{(n)}(x) < 0$ for all $x \in (a-\delta, a+\delta)$. Since n is even, $(x-a)^n \geq 0$ for all $x \in I$, hence it follows from (4.4) that

$$f(x) - f(a) \leq 0 \quad \Leftrightarrow \quad f(x) \leq f(a) \qquad \forall x \in (a-\delta, a+\delta).$$

Hence, f has a local maximum in a.

c) This follows in analogy with part b).

□

Example 4.47. We want to find the local extrema of $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 e^{-2x}$. f is a product of differentiable functions, hence itself differentiable, and by the product rule

$$f'(x) = 3x^2 e^{-2x} - 2x^3 e^{-2x} = (3x^2 - 2x^3)e^{-2x} = (3 - 2x)x^2 e^{-2x}$$

By Fermat's theorem, the only possible local extremum points are the solutions of $f'(x) = 0$, which are $x = 0$ and $x = \frac{3}{2}$. f' is again differentiable with

$$f''(x) = (6x - 6x^2)e^{-2x} - 2(3x^2 - 2x^3)e^{-2x} = (6x - 12x^2 + 4x^3)e^{-2x} = 2x(3 - 6x + 2x^2)e^{-2x}.$$

We compute that

$$f''(0) = 0, \quad f''\left(\tfrac{3}{2}\right) = 3(3 - 9 + \tfrac{9}{2})e^{-3} = -\frac{9}{2}e^{-3} < 0.$$

By Theorem 4.46, $\frac{3}{2}$ is a local maximum point and $f(\frac{3}{2}) = \frac{27}{8}e^{-3}$ is a local maximum of f.
f'' is again differentiable and we compute that

$$f'''(x) = (6 - 24x + 12x^2)e^{-2x} - 2(6x - 12x^2 + 4x^3)e^{-2x},$$

from which we see that $f'''(0) = 6e^0 = 6 \neq 0$. Hence, f has neither a local maximum nor a local minimum in 0 by Theorem 4.46.

Chapter 5

Integration

5.1 Definition of the Riemann integral

The graph of a function $f : [a, b] \to \mathbb{R}$ is formally defined as

$$\text{Graph}(f) := \{(x, f(x)) \in \mathbb{R}^2 \mid x \in [a, b]\}.$$

We want to measure the surface area of the subset of $\mathbb{R}^2$ that is enclosed by $\text{Graph}(f)$, the x-axis and the two vertical lines going through $x = a$ and through $x = b$. (See the lecture video for an illustration.)
The strategy to compute the area of this particular set is to approximate the area by considering only subsets of $\mathbb{R}^2$ whose areas are easy to compute, namely rectangles. We want to construct families of rectangles whose union is close to the original set in the first place and then study families of thinner and thinner rectangles to approximate the original set. (See again the lecture video for an illustration.) These unions of rectangles are formalized in the following definition.

Definition 5.1. Let $a, b \in \mathbb{R}$ with $a < b$.

a) A finite family of real numbers $(x_0, x_1, \ldots, x_n)$, where $n \in \mathbb{N}$ and

$$a = x_0 \leq x_1 \leq \cdots \leq x_{n-1} \leq x_n = b$$

is called *a partition of* $[a, b]$ We denote the set of all partitions of $[a, b]$ by $\mathcal{P}([a, b])$.

b) Let $P = (x_0, \ldots, x_n) \in \mathcal{P}([a, b])$. A partition $P' = (y_0, \ldots, y_m) \in \mathcal{P}([a, b])$ is called *a refinement of* P if $x_i \in \{y_0, \ldots, y_m\}$ for all $i \in \{0, 1, \ldots, n\}$. If $P_1, P_2 \in \mathcal{P}([a, b])$ are given, then $P \in \mathcal{P}([a, b])$ will be called a *common refinement of* P_1 *and* P_2 if P is both a refinement of P_1 and a refinement of P_2. (One easily checks that any two partitions of $[a, b]$ have a common refinement.)

c) Let $f : [a, b] \to \mathbb{R}$ be bounded and let $P = (x_0, x_1, \ldots, x_n) \in \mathcal{P}([a, b])$. The *upper sum of* f *over* P is given as

$$U(f, P) := \sum_{j=1}^{n} \sup_{x \in [x_{j-1}, x_j]} f(x) \cdot (x_j - x_{j-1}).$$

The *lower sum of* f *over* P is given as

$$L(f, P) := \sum_{j=1}^{n} \inf_{x \in [x_{j-1}, x_j]} f(x) \cdot (x_j - x_{j-1}).$$

(See the lecture video for an illustration.)

Theorem 5.2. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $P \in \mathcal{P}([a,b])$.*

a) $L(f,P) \leq U(f,P)$.

b) If $P' \in \mathcal{P}([a,b])$ is a refinement of P, then

$$L(f,P) \leq L(f,P') \qquad \text{and} \qquad U(f,P) \geq U(f,P').$$

Proof. a) This is obvious.

b) Let $P = (x_0, \dots, x_n)$ and $P' = (y_0, \dots, y_m)$. For each $j \in \{0, \dots, n\}$ let $k_j \in \{0, \dots, m\}$ such that $x_j = y_{k_j}$. Then

$$\begin{aligned} L(f,P) &= \sum_{j=0}^{n-1} \inf_{x \in [x_j, x_{j+1}]} f(x) \cdot (x_{j+1} - x_j) = \sum_{j=0}^{n-1} \inf_{x \in [y_{k_j}, y_{k_{j+1}}]} f(x) \cdot (y_{k_{j+1}} - y_{k_j}) \\ &= \sum_{j=0}^{n-1} \inf_{x \in [y_{k_j}, y_{k_{j+1}}]} f(x) \sum_{i=k_j}^{k_{j+1}-1} (y_{i+1} - y_i) = \sum_{j=0}^{n-1} \sum_{i=k_j}^{k_{j+1}-1} \inf_{x \in [y_{k_j}, y_{k_{j+1}}]} f(x) \cdot (y_{i+1} - y_i) \end{aligned}$$

where we used that the additional sum over i is a telescopic sum. By Exercise 1 from Sheet 2, it holds for all $j \in \{0, 1, \dots, n-1\}$ and $i \in \{k_j, k_j + 1, \dots, k_{j+1} - 1\}$ that

$$\inf_{x \in [y_{k_j}, y_{k_{j+1}}]} f(x) = \inf f([y_{k_j}, y_{k_{j+1}}]) \leq \inf f([y_i, y_{i+1}]) = \inf_{x \in [y_i, y_{i+1}]} f(x).$$

Inserting this into the above yields

$$\begin{aligned} L(f,P) &\leq \sum_{j=0}^{n-1} \sum_{i=k_j}^{k_{j+1}-1} \inf_{x \in [y_i, y_{i+1}]} f(x) \cdot (y_{i+1} - y_i) \\ &= \sum_{i=0}^{m-1} \inf_{x \in [y_i, y_{i+1}]} f(x) \cdot (y_{i+1} - y_i) = L(f,P'), \end{aligned}$$

which we wanted to show. The inequality $U(f,P) \geq U(f,P')$ is proven analogously. □

Using upper and lower sums, we next introduce the notions of upper and lower integrals which are the abovementioned approximations of the area of the surface "below a graph".

Definition 5.3. Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a,b] \to \mathbb{R}$ be bounded.

a) *The upper integral of f* is given by

$$\overline{\int_a^b} f(x)\, dx := \inf \{ U(f,P) \mid P \in \mathcal{P}([a,b]) \}.$$

The lower integral of f is given by

$$\underline{\int_a^b} f(x)\, dx := \sup \{ L(f,P) \mid P \in \mathcal{P}([a,b]) \}.$$

b) f is *Riemann-integrable*, or simply *integrable*, if

$$\overline{\int_a^b} f(x)\, dx = \underline{\int_a^b} f(x) dx.$$

In this case, we put $\int_a^b f(x)\, dx := \overline{\int_a^b} f(x) dx$ and call it *the integral of f over $[a,b]$*.

c) We put $\mathcal{R}([a,b]) := \{f : [a,b] \to \mathbb{R} \mid f \text{ is Riemann-integrable}\}$.

Example 5.4. (1) Let $f : [a,b] \to \mathbb{R}$, $f(x) = c$, be constant, where $c \in \mathbb{R}$. Then for each $P = (x_0, x_1, \ldots, x_n) \in \mathcal{P}([a,b])$ it holds that

$$U(f,P) = L(f,P) = \sum_{j=0}^{n-1} c(x_{j+1} - x_j) = c \sum_{j=0}^{n-1} (x_{j+1} - x_j) = c(b-a).$$

Thus, f is Riemann-integrable with $\int_a^b f(x) dx = c(b-a)$.

(2) Consider the function

$$f : [0,1] \to \mathbb{R}, \qquad f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

Since every non-empty open interval (a,b) contains both rational and irrational (i.e. real, but not rational) numbers, it holds that $\sup_{x \in (a,b)} f(x) = 1$ and $\inf_{x \in (a,b)} f(x) = 0$ for each such interval. From this observation, it follows that for each $P = (x_0, \ldots, x_n) \in \mathcal{P}([0,1])$

$$L(f,P) = 0, \quad U(f,P) = \sum_{j=0}^{n-1} 1 \cdot (x_j - x_{j-1}) = 1 - 0 = 1.$$

Consequently, $\underline{\int_0^1} f(x)\, dx = 0$, while $\overline{\int_0^1} f(x)\, dx = 1$, such that f is *not* Riemann-integrable.

Theorem 5.5. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a,b] \to \mathbb{R}$ be bounded. Then*

$$\underline{\int_a^b} f(x)\, dx \leq \overline{\int_a^b} f(x)\, dx.$$

Proof. Let $P_0, P_1, P_2 \in \mathcal{P}([a,b])$, such that P_0 is a common refinement of P_1 and P_2. Then by Theorem 5.2,

$$L(f,P_1) \leq L(f,P_0) \leq U(f,P_0) \leq U(f,P_2).$$

Thus, for fixed P_2, the number $U(f,P_2)$ is an upper bound of $\{L(f,P) \mid P \in \mathcal{P}([a,b])\}$, which implies

$$\underline{\int_a^b} f(x)\, dx = \sup\{L(f,P) \mid P \in \mathcal{P}([a,b])\} \leq U(f,P_2).$$

Thus, $\underline{\int_a^b} f(x)\, dx$ is a lower bound of $\{U(f,P) \mid P \in \mathcal{P}([a,b])\}$, which in turn implies

$$\underline{\int_a^b} f(x)\, dx \leq \inf\{U(f,P) \mid P \in \mathcal{P}([a,b])\} = \overline{\int_a^b} f(x)\, dx.$$

□

Example 5.6. Let $f : [a,b] \to \mathbb{R}$ be a *step function*, i.e. assume there exist $a = x_0 \leq x_1 \leq \cdots \leq xn = b$ and $c_1, \ldots, c_n \in \mathbb{R}$ with

$$f(x) = c_1 \quad \forall x \in [a, x_1], \qquad f(x) = c_{j+1} \quad \forall x \in (x_j, x_{j+1}],\ j \in \{0, 1, \ldots, n-1\}.$$

One checks that for $P = (x_0, \ldots, x_n) \in \mathcal{P}([a,b])$ it holds that

$$L(f,P) = U(f,P) = \sum_{j=0}^{n-1} c_{j+1}(x_{j+1} - x_j).$$

Thus, by Theorem 5.5,

$$\sum_{j=0}^{n-1} c_{j+1}(x_{j+1} - x_j) \leq \underline{\int_a^b} f(x)\, dx \leq \overline{\int_a^b} f(x)\, dx = \sum_{j=0}^{n-1} c_{j+1}(x_{j+1} - x_j).$$

Hence, f is Riemann-integrable with $\int_a^b f(x)\, dx = \sum_{j=0}^{n-1} c_{j+1}(x_{j+1} - x_j)$.

The following theorem provides a tangible criterion on a function to be Riemann-integrable.

Theorem 5.7. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a,b] \to \mathbb{R}$ be bounded. f is Riemann-integrable if and only if*

$$\forall \varepsilon > 0\ \exists P \in \mathcal{P}([a,b]) : \quad U(f,P) - L(f,P) < \varepsilon.$$

Proof. "⇒": Let $\varepsilon > 0$. Since f is Riemann-integrable, there are $P_1, P_2 \in \mathcal{P}([a,b])$ with

$$\int_a^b f(x)dx - L(f,P_1) < \frac{\varepsilon}{2}, \qquad U(f,P_2) - \int_a^b f(x) < \frac{\varepsilon}{2}.$$

Let P_0 be a common refinement of P_1 and P_2. Then, by Theorem 5.2,

$$U(f,P_0) \leq U(f,P_1) < \int_a^b f(x)dx + \frac{\varepsilon}{2} < L(f,P_1) + \varepsilon \leq L(f,P_0) + \varepsilon,$$

which shows the claim.
"⇐": By construction and by Theorem 5.5, it holds for each $P \in \mathcal{P}([a,b])$ that

$$L(f,P) \leq \underline{\int_a^b} f(x)dx \leq \overline{\int_a^b} f(x)dx \leq U(f,P).$$

Given $\varepsilon > 0$ and $P \in \mathcal{P}([a,b])$ with $U(f,P) - L(f,P) < \varepsilon$, we derive

$$0 \leq \overline{\int_a^b} f(x)dx - \underline{\int_a^b} f(x)dx \leq U(f,P) - L(f,P) < \varepsilon.$$

Since this shows that $0 \leq \overline{\int_a^b} f(x)dx - \underline{\int_a^b} f(x)dx < \varepsilon$ for all $\varepsilon > 0$, it follows that $\underline{\int_a^b} f(x)dx = \overline{\int_a^b} f(x)dx$. □

Let's apply this criterion to a specific example.

Example 5.8. Let $b > 0$, and consider $f : [0,b] \to \mathbb{R}$, $f(x) = x$. For each $n \in \mathbb{N}$ let $P_n = (x_0, \dots, x_n) \in \mathcal{P}([0,b])$ be given by $x_j = \frac{j}{n} b$ for each $j \in \{0, 1, \dots, n\}$. We compute the lower and upper sums of f with respect to P_n as

$$\begin{aligned} U(f, P_n) &= \sum_{j=0}^{n-1} \sup_{x \in [\frac{j}{n}b, \frac{j+1}{n}b]} x \cdot \left(\tfrac{j+1}{n} b - \tfrac{j}{n} b \right) \\ &= \sum_{j=0}^{n-1} \frac{j+1}{n} b \cdot \frac{b}{n} = \sum_{j=0}^{n-1} (j+1) \frac{b^2}{n^2} = \frac{n(n+1)}{2} \cdot \frac{b^2}{n^2} = \frac{(1+\frac{1}{n})b^2}{2}, \\ L(f, P_n) &= \sum_{j=0}^{n-1} \frac{j}{n} b \cdot \frac{b}{n} = \sum_{j=0}^{n-1} j \frac{b^2}{n^2} = \frac{n(n-1)}{2} \cdot \frac{b^2}{n^2} = \frac{(1-\frac{1}{n})b^2}{2}. \end{aligned}$$

where we have used Theorem 1.62. We derive that

$$U(f, P_n) - L(f, P_n) = \frac{\frac{2}{n} b^2}{2} = \frac{b^2}{n} \qquad \forall n \in \mathbb{N},$$

hence $\lim_{n\to\infty}(U(f, P_n) - L(f, P_n)) = 0$. From this observation and Theorem 5.7, one derives that f is Riemann-integrable and that

$$\int_0^b x \, dx = \lim_{n\to\infty} L(f, P_n) = \lim_{n\to\infty} U(f, P_n) = \lim_{n\to\infty} \frac{(1+\frac{1}{n})b^2}{2} = \frac{b^2}{2}.$$

Using the criterion from Theorem 5.7, we will next show that two big classes of functions are Riemann-integrable.

Theorem 5.9. *Let $a, b \in \mathbb{R}$ with $a < b$. Every continuous function $f : [a,b] \to \mathbb{R}$ is Riemann-integrable.*

Proof. Let $\varepsilon > 0$. By Theorem 3.38, f is uniformly continuous, hence there exists $\delta > 0$, such that

$$|f(x) - f(y)| < \frac{\varepsilon}{b-a} \quad \forall x, y \in [a,b] \text{ with } |x - y| < \delta. \tag{5.1}$$

Let $P = (x_0, x_1, \dots, x_n) \in \mathcal{P}([a,b])$ be given, such that $x_{j+1} - x_j < \delta$ for all $j \in \{0, 1, \dots, n-1\}$. Then by Theorem 3.32,

$$\sup_{x \in [x_j, x_{j+1}]} f(x) - \inf_{x \in [x_j, x_{j+1}]} f(x) = \max_{x \in [x_j, x_{j+1}]} f(x) - \min_{x \in [x_j, x_{j+1}]} f(x) \overset{(5.1)}{<} \frac{\varepsilon}{b-a}$$

for all $j \in \{0, 1, \dots, n-1\}$. Consequently,

$$\begin{aligned} U(f, P) - L(f, P) &= \sum_{j=0}^{n-1} \Big(\sup_{x \in [x_j, x_{j+1}]} f(x) - \inf_{x \in [x_j, x_{j+1}]} f(x) \Big) (x_{j+1} - x_j) \\ &< \frac{\varepsilon}{b-a} \sum_{j=0}^{n-1} (x_{j+1} - x_j) = \frac{\varepsilon}{b-a} (x_n - x_0) = \frac{\varepsilon}{b-a} (b-a) = \varepsilon. \end{aligned}$$

Thus, f is Riemann-integrable by Theorem 5.7. □

Theorem 5.10. *Let $a, b \in \mathbb{R}$ with $a < b$. Every monotonic function $f : [a, b] \to \mathbb{R}$ is Riemann-integrable.*

Proof. (This proof is not discussed in the lecture videos.)

Assume throughout the proof that f is monotonically increasing, the monotonically decreasing case is proven along the same lines. If $f(a) = f(b)$, then f is constant and the statement was shown in Example 5.4, hence assume that $f(a) < f(b)$.
Let $\varepsilon > 0$. Given $n \in \mathbb{N}$ we define a partition $P_n = (x_0, x_1, \dots, x_n)$ by

$$x_j = a + \frac{j}{n}(b-a) \qquad \forall j \in \{0, 1, \dots, n\},$$

such that $x_{j+1} - x_j = \frac{b-a}{n}$ for all $j \in \{0, 1, \dots, n-1\}$. Since f is monotonically increasing,

$$\sup_{x \in [x_j, x_{j+1}]} f(x) = f(x_{j+1}), \qquad \inf_{x \in [x_j, x_{j+1}]} f(x) = f(x_j) \quad \forall j \in \{0, 1, \dots, n-1\},$$

such that

$$U(f, P_n) = \sum_{j=0}^{n-1} f(x_{j+1})(x_{j+1} - x_j) = \sum_{j=0}^{n-1} f(x_{j+1}) \frac{b-a}{n},$$
$$L(f, P_n) = \sum_{j=0}^{n-1} f(x_j)(x_{j+1} - x_j) = \sum_{j=0}^{n-1} f(x_j) \frac{b-a}{n}.$$

Hence,

$$\begin{aligned} U(f, P_n) - L(f, P_n) &= \Big(\sum_{j=0}^{n-1} f(x_{j+1}) - \sum_{j=0}^{n-1} f(x_j) \Big) \frac{b-a}{n} \\ &= (f(x_n) - f(x_0)) \frac{b-a}{n} = \frac{(f(b) - f(a))(b-a)}{n}. \end{aligned}$$

Thus, if $n \in \mathbb{N}$ is chosen so big that $\frac{1}{n} < \frac{\varepsilon}{(f(b)-f(a))(b-a)}$, then $U(f, P_n) - L(f, P_n) < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, the claim follows from Theorem 5.7. □

5.2 Properties of the Riemann integral

Having established the definition of the Riemann integral and having shown the integrability of large classes of functions, we want to study basic properties of the Riemann integral, like its compatibility with algebraic operations.
The following theorem summarizes the most important computational properties of the Riemann integral that will be frequently used in the rest of this chapter.

Theorem 5.11. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f, g \in \mathcal{R}([a, b])$. Then:*

a) $f + g \in \mathcal{R}([a, b])$ and $\int_a^b (f+g)(x)\, dx = \int_a^b f(x)\, dx + \int_a^b g(x)\, dx$.

b) For all $\lambda \in \mathbb{R}$ it holds that $\lambda f \in \mathcal{R}([a, b])$ and $\int_a^b (\lambda f)(x)\, dx = \lambda \int_a^b f(x)\, dx$.

c) If $f(x) \leq g(x)$ for all $x \in [a,b]$, then $\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx$.

d) If $c \in (a,b)$, then $f|_{[a,c]} \in \mathcal{R}([a,c])$, $f|_{[c,b]} \in \mathcal{R}([c,b])$ and

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

e) If $m, M \in \mathbb{R}$ satisfy $m \leq f(x) \leq M$ for all $x \in [a,b]$, then

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a).$$

Proof. (The proof of parts b) to e) are not discussed in the lecture videos.)

a) Let $P = (x_0, x_1, \ldots, x_n) \in \mathcal{P}([a,b])$. Then for each $j \in \{0,1,\ldots,n-1\}$

$$\sup_{x\in[x_j,x_{j+1}]} (f+g)(x) = \sup_{x\in[x_j,x_{j+1}]} (f(x)+g(x)) \leq \sup_{x\in[x_j,x_{j+1}]} f(x) + \sup_{y\in[x_j,x_{j+1}]} g(y)$$

and analogously,

$$\inf_{x\in[x_j,x_{j+1}]} (f+g)(x) \geq \inf_{x\in[x_j,x_{j+1}]} f(x) + \inf_{y\in[x_j,x_{j+1}]} g(y).$$

From these inequalities, one straightforwardly derives that

$$U(f+g,P) \leq U(f,P) + U(g,P), \qquad L(f+g,P) \geq L(f,P) + L(g,P).$$

Let $\varepsilon > 0$. Since f and g are Riemann-integrable, by Theorem 5.7 there exist $P_1, P_2 \in \mathcal{P}([a,b])$, such that $U(f,P_1) - L(f,P_1) < \frac{\varepsilon}{2}$ and $U(g,P_2) - L(g,P_2) < \frac{\varepsilon}{2}$. Let $P_0 \in \mathcal{P}([a,b])$ be a common refinement of P_1 and P_2. By the previous inequalities and by Theorem 5.2,

$$\begin{aligned} U(f+g,P_0) - L(f+g,P_0) &\leq U(f,P_0) + U(g,P_0) - L(f,P_0) - L(g,P_0) \\ &\leq U(f,P_1) - L(f,P_1) + U(g,P_2) - L(g,P_2) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Hence, $f+g$ is Riemann-integrable by Theorem 5.7. By choice of P_1 and P_2 it holds that $U(f,P_0) \leq U(f,P_1) < \int_a^b f(x)\,dx + \frac{\varepsilon}{2}$ and $U(g,P_0) < \int_a^b g(x)\,dx + \frac{\varepsilon}{2}$ and thus

$$\int_a^b (f+g)(x)\,dx \leq U(f+g,P_0) \leq U(f,P_0) + U(g,P_0) < \int_a^b f(x)\,dx + \int_a^b g(x)\,dx + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that

$$\int_a^b (f+g)(x)\,dx \leq \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$

The opposite inequality is derived analogously, using that $L(f,P_0) > \int_a^b f(x)\,dx - \frac{\varepsilon}{2}$ and $L(g,P_0) > \int_a^b g(x)\,dx - \frac{\varepsilon}{2}$.

b) Let $\lambda \in \mathbb{R}$. As one easily sees, it holds that

$$U(\lambda f, P) = \lambda U(f, P), \quad L(\lambda f, P) = \lambda L(f, P) \quad \forall P \in \mathcal{P}([a, b]).$$

From this observation and the properties of supremum and infimum, one derives that

$$\overline{\int_a^b} \lambda f(x)\, dx = \inf\{\lambda U(f, P) \mid P \in \mathcal{P}([a, b])\} = \lambda \inf\{U(f, P) \mid P \in \mathcal{P}([a, b])\} = \lambda \overline{\int_a^b} f(x)\, dx$$

and analogously, $\underline{\int_a^b} \lambda f(x)\, dx = \lambda \underline{\int_a^b} f(x)\, dx$. Consequently, since f is Riemann-integrable, $\underline{\int_a^b} \lambda f(x)\, dx = \overline{\int_a^b} \lambda f(x)\, dx = \lambda \int_a^b f(x)\, dx$.

c) If $f(x) \leq g(x)$ for all $x \in [a, b]$, then for each $P = (x_0, \ldots, x_n) \in \mathcal{P}([a, b])$ we compute that

$$U(f, P) = \sum_{j=0}^{n-1} \sup_{x \in [x_j, x_{j+1}]} f(x) \cdot (x_{j+1} - x_j) \leq \sum_{j=0}^{n-1} \sup_{x \in [x_j, x_{j+1}]} g(x) \cdot (x_{j+1} - x_j) = U(g, P).$$

From this observation and the properties of the infimum, we derive

$$\int_a^b f(x)\, dx = \inf\{U(f, P) \mid P \in \mathcal{P}([a, b])\} \leq \inf\{U(g, P) \mid P \in \mathcal{P}([a, b])\} = \int_a^b g(x)\, dx.$$

d) If $P_1 = (x_0, x_1, \ldots, x_n) \in \mathcal{P}([a, c])$ and $P_2 = (y_0, y_1, \ldots, y_m) \in \mathcal{P}([c, b])$. Then $P_0 := (x_0, \ldots, x_n, y_1, \ldots, y_m) \in \mathcal{P}([a, b])$ and it obviously holds that

$$U(f|_{[a,c]}, P_1) + U(f|_{[c.b]}, P_2) = U(f, P_0), \qquad L(f|_{[a,c]}, P_1) + L(f|_{[c.b]}, P_2) = L(f, P_0).$$

Let $\mathcal{P}' = \{P = (x_0, \ldots, x_n) \in \mathcal{P}([a, b]) \mid \exists k \in \{0, 1, \ldots, n\} \text{ with } x_k = c\}$.
Let $\varepsilon > 0$. Since f is Riemann-integrable, there exists $P \in \mathcal{P}([a, b])$ with $U(f, P) - L(f, P) < \varepsilon$. Let P' be a refinement of P with $P' \in \mathcal{P}'$. Then $U(f, P') - L(f, P') < \varepsilon$ as well.
If $P' = (x_0, \ldots, x_n)$ with $x_k = c$, it holds that

$$P_1 := (x_0, \ldots, x_k) \in \mathcal{P}([a, c]) \quad \text{and} \quad P_2 := (x_k, \ldots, x_n) \in \mathcal{P}([c, b])$$

and one easily checks that $U(f, P) = U(f|_{[a,c]}, P_1) + U(f|_{[c,b]}, P_2)$ and $L(f, P) = L(f|_{[a,c]}, P_1) + L(f|_{[c,b]}, P_2)$. Thus,

$$U(f|_{[a,c]}, P_1) - L(f|_{[a,c]}, P_1) + U(f|_{[c,b]}, P_2) - L(f|_{[c,b]}, P_2) < \varepsilon$$

and since the two differences are non-negative, this yields $U(f|_{[a,c]}, P_1) - L(f|_{[a,c]}, P_1) < \varepsilon$ and $U(f|_{[c,b]}, P_2) - L(f|_{[c,b]}, P_2) < \varepsilon$. Since ε was arbitrary, this shows that $f|_{[a,c]}$ and $f|_{[c,b]}$ are Riemann-integrable and a simple computation that is left to the reader shows the claimed formula for the corresponding integrals.

e) Let $P = (x_0, \ldots, x_n) \in \mathcal{P}([a, b])$. Then

$$U(f, P) = \sum_{j=0}^{n-1} \sup_{x \in [x_j, x_{j+1}]} f(x) \cdot (x_{j+1} - x_j) \leq M \sum_{j=0}^{n-1} (x_{j+1} - x_j) = M(b - a),$$

$$L(f, P) = \sum_{j=0}^{n-1} \inf_{x \in [x_j, x_{j+1}]} f(x) \cdot (x_{j+1} - x_j) \geq m \sum_{j=0}^{n-1} (x_{j+1} - x_j) = m(b - a).$$

Thus, by definition of the Riemann integral,

$$m(b-a) \leq L(f,P) \leq \int_a^b f(x)\,dx \leq U(f,P) \leq M(b-a).$$

□

In contrast to the compatibility of continuity and differentiability with the composition of functions, the composition of two Riemann-integrable functions does *not* necessarily have to be Riemann-integrable. Such "non-examples" are usually not very intuitive and we refrain from giving an explicit example. Instead we prove that a slightly stricter assumption ensures the Riemann-integrability of a composition.

Theorem 5.12. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $I \subset \mathbb{R}$ be an interval. If $f : [a,b] \to I$ is Riemann-integrable and $g : I \to \mathbb{R}$ is continuous, then $g \circ f : [a,b] \to \mathbb{R}$ is Riemann-integrable.*

Proof. Since f is Riemann-integrable, it is bounded and we can assume w.l.o.g. that I is compact. Let $\varepsilon > 0$. Since I is compact and g is continuous, g is uniformly continuous by Theorem 3.38. Hence, there exists $\delta > 0$, such that

$$|g(x) - g(y)| < \frac{\varepsilon}{2(b-a)} \quad \forall x, y \in I \text{ with } |x-y| < \delta. \tag{5.2}$$

g is also bounded by Theorem 3.7 and we put $C := \sup_{x \in I} |g(x)|$. Since the claim is obvious for $g = 0$, we assume that $C > 0$. We further assume w.l.o.g. that δ is chosen so small that

$$\delta \leq \frac{\varepsilon}{4C}. \tag{5.3}$$

Since f is Riemann-integrable, there exists $P = (x_0, \ldots, x_n) \in \mathcal{P}([a,b])$, such that

$$U(f,P) - L(f,P) < \delta^2.$$

Put

$$\begin{aligned}
&M_j := \sup_{x \in [x_j, x_{j+1}]} f(x), \qquad M_j^* := \sup_{x \in [x_j, x_{j+1}]} (g \circ f)(x), \\
&m_j := \inf_{x \in [x_j, x_{j+1}]} f(x), \qquad m_j^* := \inf_{x \in [x_j, x_{j+1}]} (g \circ f)(x) \qquad \forall j \in \{0, 1, \ldots, n-1\}, \\
&A := \{j \in \{0, 1, \ldots, n-1\} \mid M_j - m_j < \delta\}, \qquad B := \{0, 1, \ldots, n-1\} \setminus A.
\end{aligned}$$

For each $j \in A$ we compute that

$$M_j^* - m_j^* \leq \sup_{x,y \in [x_j, x_{j+1}]} (g(f(x)) - g(f(y))) \overset{(5.2)}{\leq} \frac{\varepsilon}{2(b-a)}.$$

Hence,

$$\sum_{j \in A} (M_j^* - m_j^*)(x_{j+1} - x_j) \leq \frac{\varepsilon}{2(b-a)} \sum_{j \in A} (x_{j+1} - x_j) \leq \frac{\varepsilon}{2(b-a)} \sum_{j=0}^{n-1} (x_{j+1} - x_j) = \frac{\varepsilon}{2}.$$

For $j \in B$ it holds that $M_j - m_j \geq \delta$. Thus,

$$\delta \sum_{j \in B} (x_{j+1} - x_j) \leq \sum_{j \in B} (M_j - m_j)(x_{j+1} - x_j) \leq U(f,P) - L(f,P) < \delta^2,$$

which shows that $\sum_{j\in B}(x_{j+1}-x_j)<\delta$. Using this inequality and that

$$M_j^*-m_j^*\le \sup_{x,y\in[x_j,x_{j+1}]}|g(f(x))-g(f(y))|\le \sup_{x,y\in[x_j,x_{j+1}]}(|g(f(x))|+|g(f(y))|)\le 2C$$

holds for each j, we derive that

$$\sum_{j\in B}(M_j^*-m_j^*)(x_{j+1}-x_j)\le 2C\sum_{j\in B}(x_{j+1}-x_j)<2C\delta \overset{(5.3)}{\le} \frac{\varepsilon}{2},$$

Thus,

$$\begin{aligned}U(g\circ f,P)-L(g\circ f,P)&=\sum_{j=0}^{n-1}(M_j^*-m_j^*)(x_{j+1}-x_j)\\&=\sum_{j\in A}(M_j^*-m_j^*)(x_{j+1}-x_j)+\sum_{j\in B}(M_j^*-m_j^*)(x_{j+1}-x_j)<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.\end{aligned}$$

Since ε was chosen arbitrarily, it follows from Theorem 5.7 that $g\circ f$ is Riemann-integrable. □

Theorem 5.13. *Let $a,b\in\mathbb{R}$ with $a<b$ and let $f,g\in\mathcal{R}([a,b])$. Then:*

a) $f\cdot g\in\mathcal{R}([a,b])$.

b) $|f|\in\mathcal{R}([a,b])$ *and* $\left|\int_a^b f(x)dx\right|\le\int_a^b|f(x)|dx$.

Proof. a) In the case of $f=g$, i.e. for $f^2=f\cdot f:[a,b]\to\mathbb{R}$, the claim follows from Theorem 5.12, since $f^2=\phi\circ f$, where $\phi:\mathbb{R}\to\mathbb{R}$, $\phi(x)=x^2$.

In the general case, we observe that $f\cdot g=\frac{1}{2}((f+g)^2-f^2-g^2)$ and derive from the previous observation and Theorem 5.11.a) and b) that $f\cdot g$ is Riemann-integrable.

b) Since $|f|=\phi\circ f$, where $\phi(x)=|x|$, it follows from Theorem 5.12 that $|f|$ is Riemann-integrable. Let

$$c:=\begin{cases}1 & \text{if } \int_a^b f(x)dx\ge 0,\\ -1 & \text{if } \int_a^b f(x)dx<0.\end{cases}$$

Then by Theorem 5.11.b),

$$\left|\int_a^b f(x)\,dx\right|=c\int_a^b f(x)\,dx=\int_a^b cf(x)\,dx.$$

Since $cf(x)\le|f(x)|$ for each $x\in[a,b]$, it follows from Theorem 5.11.e) that $|\int_a^b f(x)\,dx|\le\int_a^b|f(x)|\,dx$.

□

Remark 5.14. Note that while the product of two Riemann-integrable functions is again Riemann-integrable by Theorem 5.13.a), it is not clear how the different integrals are related to each other. Later in this chapter, we will study the technique of integration by parts, which will enable us to compute the integrals of product in certain situations.

The following theorem is similar in spirit to the mean value theorem for derivatives that we have seen in Section 4.3.

Theorem 5.15 (Mean value theorem for integrals)**.** *Let $a, b \in \mathbb{R}$ with $a < b$. Let $\varphi \in \mathcal{R}([a,b])$ with $\varphi(x) \geq 0$ for all $x \in [a,b]$. Then for each continuous function $f : [a,b] \to \mathbb{R}$ there exists $\xi \in [a,b]$ with*

$$\int_a^b f(x)\varphi(x)dx = f(\xi)\int_a^b \varphi(x)dx.$$

Proof. By Theorems 5.9 and Theorem 5.13.a), $f \cdot \varphi$ is Riemann-integrable. Since f is continuous and $[a,b]$ is compact, f has a minimum and a maximum by Theorem 3.32 and we put

$$m := \min_{x\in[a,b]} f(x), \qquad M := \max_{x\in[a,b]} f(x).$$

Then by Theorem 5.11.b) and e)

$$m\int_a^b \varphi(x)\,dx = \int_a^b m\varphi(x)\,dx \leq \int_a^b f(x)\varphi(x)\,dx \leq \int_a^b M\varphi(x)\,dx = M\int_a^b \varphi(x)\,dxx.$$

Thus, there exists $\mu \in [m,M]$, such that $\int_a^b f(x)\varphi(x)\,dx = \mu\int_a^b \varphi(x)\,dx$. But by the intermediate value theorem, there exists $\xi \in [a,b]$ with $f(\xi) = \mu$, which shows the claim. □

Corollary 5.16. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a,b] \to \mathbb{R}$ be continuous. There exists $\xi \in [a,b]$ with*

$$\int_a^b f(x)dx = f(\xi)(b-a).$$

Proof. This is Theorem 5.15 in the special case of $\varphi(x) = 1$ for all $x \in [a,b]$. □

Remark 5.17. Given $f \in \mathcal{R}([a,b])$, the quantity $M(f) := \frac{1}{b-a}\int_a^b f(x)\,dx$ is called *the mean of f*. It is interpreted as the *average value* that f takes on $[a,b]$. For a continuous function f, Corollary 5.16 says that this average value is always contained in the set of values of f.
Likewise, given $\varphi \in \mathcal{R}([a,b])$ with $\varphi(x) \geq 0$ for all $x \in [a,b]$, the number

$$M_\varphi(f) := \frac{\int_a^b f(x)\varphi(x)\,dx}{\int_a^b \varphi(x)\,dx}$$

is called *the weighted mean of f with respect to φ*. Here, φ is interpreted as a function prioritizing certain parts of the interval, e.g. indicating a distribution of probabilities. Theorem 5.15 says that for continuous f every weighted average of f is among the values of f.

5.3 Integration and differentiation

Throughout this section, let I be an interval containing more than one point.

Definition 5.18. Let $a, b \in \mathbb{R}$ with $a < b$ and let $f \in \mathcal{R}([a,b])$. We put

$$\int_b^a f(x)\,dx := -\int_a^b f(x)\,dx.$$

For each $c \in [a,b]$ we further let $\int_c^c f(x)\,dx := 0$.

One checks using Theorem 5.11.d) that with this convention

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

holds for all $a, b, c \in I$.
So far, the constructions of derivatives and of integrals have been completely independent of each other. However, they are intimately connected as the following important theorem shows, which, in the lecturer's opinion, is one of the most beautiful theorems in mathematics.

Theorem 5.19 (Fundamental theorem of calculus)**.** *Let $a \in I$ and let $f : I \to \mathbb{R}$ be continuous.*

a) The function

$$F : I \to \mathbb{R}, \qquad F(x) = \int_a^x f(t)\,dt,$$

is differentiable with $F'(x) = f(x)$ for each $x \in I$.

b) Let $F : I \to \mathbb{R}$ be an arbitrary differentiable function with $F' = f$. Then

$$\int_a^b f(x)\,dx = F(b) - F(a) \qquad \forall a, b \in I.$$

Proof. a) Let $x \in I$. For $h \neq 0$ with $x + h \in I$ we compute that

$$\frac{F(x+h) - F(x)}{h} = \frac{\int_a^{x+h} f(t)\,dt - \int_a^x f(t)\,dt}{h} = \frac{1}{h}\int_x^{x+h} f(t)\,dt.$$

By Corollary 5.16 there exists $\xi_h \in [x, x+h] \cup [x+h, x]$ with

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h} f(\xi_h)(x + h - x) = f(\xi_h). \tag{5.4}$$

Now let $(h_n)_{n\in\mathbb{N}}$ be a sequence with $x + h_n \in I$ for all $n \in \mathbb{N}$ and with $\lim_{n\to\infty} h_n = 0$. For each $n \in \mathbb{N}$ let $\xi_n := \xi_{h_n}$ be given as in (5.4). Since $(h_n)_{n\in\mathbb{N}}$ converges to 0, it follows that $\lim_{n\to\infty} \xi_n = x$. Thus, since f is continuous,

$$\lim_{n\to\infty} \frac{F(x+h_n) - F(x)}{h_n} = \lim_{n\to\infty} f(\xi_n) = f(x).$$

Since $(h_n)_{n\in\mathbb{N}}$ was arbitrary, this shows that

$$F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = f(x).$$

b) Let $F_0 : I \to \mathbb{R}$, $F_0(x) = \int_a^x f(t)\,dt$. By a), $F_0'(x) = f(x)$ for all $x \in I$. By Corollary 4.25.b), there exists $c \in \mathbb{R}$ with $F(x) = F_0(x) + c$ for all $x \in I$. Hence,

$$\begin{aligned} F(b) - F(a) &= F_0(b) + c - (F_0(a) + c) = F_0(b) - F_0(a) \\ &= \int_a^b f(x)\,dx - \int_a^a f(x)\,dx = \int_a^b f(x)\,dx. \end{aligned}$$

□

Definition 5.20. An *antiderivative of* $f : I \to \mathbb{R}$ is a differentiable function $F : I \to \mathbb{R}$ with $F'(x) = f(x)$ for all $x \in I$. An antiderivative of f is also denoted by $\int f(x)dx$.

By part a) of the fundamental theorem of calculus, every continuous function $f : I \to \mathbb{R}$ possesses an antiderivative. In the following we will always denote

$$F(x)|_a^b := F(b) - F(a).$$

Example 5.21. Our computations of derivatives in Chapter 4 give rise to some explicit antiderivatives, which in turn allows for the computation of integrals using the fundamental theorem of calculus. Let $a, b \in \mathbb{R}$ with $a < b$.

- For $r \in \mathbb{R} \setminus \{-1\}$ it holds that $\displaystyle\int_a^b x^r \, dx = \frac{x^{r+1}}{r+1}\Big|_a^b$, where in the case of $r < 0$ we assume that $a > 0$ or $b < 0$.
- If $a > 0$, then $\int_a^b \frac{1}{x} \, dx = \log x\Big|_a^b$. If $b < 0$, then $\int_a^b \frac{1}{x} \, dx = -\int_a^b \frac{1}{-x} dx = -\left(-\log(-x)\right)\Big|_a^b = \log(-x)\Big|_a^b$. Both cases are summarized by

$$\int_a^b \frac{1}{x} \, dx = \log(|x|)\Big|_a^b \quad \text{if } 0 \notin [a, b].$$

- $\displaystyle\int_a^b e^{cx} \, dx = \frac{e^{cx}}{c}\Big|_a^b \qquad \forall c \in \mathbb{R} \setminus \{0\}.$
- $\displaystyle\int_a^b \sin x \, dx = -\cos x\Big|_a^b$ and $\displaystyle\int_a^b \cos x \, dx = \sin x\Big|_a^b$.
- If $[a, b] \subset (-\frac{\pi}{2}, \frac{\pi}{2})$, then $\displaystyle\int_a^b \frac{1}{\cos^2(x)} \, dx = \tan x\Big|_a^b$.
- $\displaystyle\int_a^b \frac{1}{1+x^2} \, dx = \arctan x\Big|_a^b$.
- If $[a, b] \subset (-1, 1)$, then $\displaystyle\int_a^b \frac{1}{\sqrt{1-x^2}} \, dx = \arcsin x\Big|_a^b$.

Our next strategy is to find formulas for integrals of more complicated functions by "inverting" some rules of differentiation that we have established in Chapter 4, more precisely the product rule and the chain rule. Before doing so, we introduce one more bit of terminology.

Definition 5.22. We call $f : I \to \mathbb{C}$ *continuously differentiable* (or *of class* C^1) if f is differentiable and its derivative $f' : I \to \mathbb{C}$ is continuous.

Theorem 5.23 (Integration by parts)**.** *Let $a, b \in \mathbb{R}$ with $a < b$. Let $u, v : [a, b] \to \mathbb{R}$ be continuously differentiable. Then*

$$\int_a^b u'(x)v(x) \, dx = u(x)v(x)\Big|_a^b - \int_a^b u(x)v'(x) \, dx.$$

Proof. By the product rule, $u \cdot v$ is again differentiable with

$$(u \cdot v)'(x) = u'(x)v(x) + u(x)v'(x) \quad \Leftrightarrow \quad u'(x)v(x) = (u \cdot v)'(x) - u'(x)v(x).$$

Since $(u \cdot v)'$ is again continuous, it follows from the fundamental theorem of calculus that

$$\begin{aligned}\int_a^b u'(x)v(x)\,dx &= \int_a^b ((u\cdot v)'(x) - u(x)v'(x))\,dx \\ &= \int_a^b (u\cdot v)'(x)\,dx - \int_a^b u(x)v'(x)\,dx = u(x)v(x)\Big|_a^b - \int_a^b u(x)v'(x)\,dx.\end{aligned}$$

□

Example 5.24. (1) Let $0 < a < b$. We want to compute $\int_a^b \log x\,dx$. For this, we apply integration by parts to the functions $u(x) = x$, $v(x) = \log x$, which satisfy $u'(x) = 1$ and $v'(x) = \frac{1}{x}$. Thus, integration by parts yields

$$\begin{aligned}\int_a^b \log x\,dx &= \int_a^b 1\cdot\log x\,dx = x\log x|_a^b - \int_a^b x\cdot\frac{1}{x}\,dx \\ &= x\log x|_a^b - \int_a^b 1\,dx = (x\log x - x)\Big|_a^b = b\log b - b - a\log a + a.\end{aligned}$$

(2) Let $c \in \mathbb{R}\setminus\{0\}$. We want to compute $\int_a^b e^{cx}\sin x\,dx$. We apply integration by parts with $u(x) = \frac{1}{c}e^{cx}$, $v(x) = \sin x$, such that $u'(x) = e^{cx}$, $v'(x) = \cos x$ for all $x \in \mathbb{R}$, and obtain:

$$\int_a^b e^{cx}\sin x dx = \frac{e^{cx}\sin x}{c}\Big|_a^b - \frac{1}{c}\int_a^b e^{cx}\cos x\,dx.$$

The integral on the right-hand side is computed analogously with $u(x) = \frac{1}{c}e^{cx}$ and $v(x) = \cos x$ as

$$\int_a^b e^{cx}\cos x = \frac{e^{cx}\cos x}{c}\Big|_a^b + \frac{1}{c}\int_a^b e^{cx}\sin x\,dx.$$

Inserting this into the above yields

$$\begin{aligned}&\int_a^b e^{cx}\sin x dx = \frac{e^c x\sin x}{c} - \frac{e^{cx}\cos x}{c^2}\Big|_a^b - \frac{1}{c^2}\int_a^b e^{cx}\sin x dx \\ \Leftrightarrow\quad &\left(1+\frac{1}{c^2}\right)\int_a^b e^{cx}\sin x\,dx = \frac{e^{cx}(c\sin x - \cos x)}{c^2}\Big|_a^b \\ \Leftrightarrow\quad &\int_a^b e^{cx}\sin x\,dx = \frac{e^{cx}(c\sin x - \cos x)}{1+c^2}\Big|_a^b.\end{aligned}$$

In the following theorem, we use the chain rule "backwards" to obtain a result about integrals.

Theorem 5.25 (Integration by substitution)**.** *Let $a, b \in \mathbb{R}$ with $a < b$. Let $f : I \to \mathbb{R}$ be continuous and let $\varphi : [a,b] \to \mathbb{R}$ be continuously differentiable with $\varphi([a,b]) \subset I$. Then:*

$$\int_a^b f(\varphi(t))\varphi'(t)\,dt = \int_{\varphi(a)}^{\varphi(b)} f(x)\,dx.$$

Proof. Let $F : I \to \mathbb{R}$ be an antiderivative of f, which exists by the fundamental theorem of calculus. Then $F\circ\varphi : [a,b]$ is continuously differentiable and by the chain rule,

$$(F\circ\varphi)'(t) = F'(\varphi(t))\varphi'(t) \quad \forall t \in [a,b].$$

Thus, by the fundamental theorem of calculus,

$$\int_a^b F'(\varphi(t))\varphi'(t)\,dt = \int_a^b (F\circ\varphi)'(t)\,dt = (F\circ\varphi)(b) - (F\circ\varphi)(a)$$
$$= F(\varphi(b)) - F(\varphi(a)) = \int_{\varphi(a)}^{\varphi(b)} f(x)\,dx.$$

□

Corollary 5.26. *Let $a, b \in \mathbb{R}$ with $a < b$.*

a) Let $c \in \mathbb{R}$ and let $f : [a+c, b+c] \to \mathbb{R}$ be continuous. Then

$$\int_a^b f(t+c)\,dt = \int_{a+c}^{b+c} f(x)\,dx.$$

b) Let $c \in \mathbb{R} \setminus \{0\}$ and let $f : [ca, cb] \to \mathbb{R}$ be continuous. Then

$$\int_a^b f(ct)\,dt = \frac{1}{c}\int_{ca}^{cb} f(x)\,dx.$$

c) Let $f : [a^2, b^2] \to \mathbb{R}$ be continuous. Then

$$\int_a^b tf(t^2)\,dt = \frac{1}{2}\int_{a^2}^{b^2} f(x)\,dx.$$

d) Let $\varphi : [a, b] \to \mathbb{R}$ be continuously differentiable with $\varphi(t) \neq 0$ for all $t \in [a, b]$. Then

$$\int_a^b \frac{\varphi'(t)}{\varphi(t)}\,dt = \log(|\varphi(t)|)\Big|_a^b.$$

Proof. a) Apply Theorem 5.25 with $\varphi(t) = t + c$, hence $\varphi'(t) = 1$ for all $t \in \mathbb{R}$.

b) Apply Theorem 5.25 with $\varphi(t) = ct$, hence $\varphi'(t) = c$ for all $t \in \mathbb{R}$ and thus

$$\int_a^b f(ct)\,dt = \frac{1}{c}\int_a^b f(ct)c\,dt = \frac{1}{c}\int_a^b f(\varphi(t))\varphi'(t)\,dt = \frac{1}{c}\int_a^b f(x)\,dx.$$

c) Apply Theorem 5.25 with $\varphi(t) = t^2$, hence $\varphi'(t) = 2t$ for all $t \in \mathbb{R}$.

d) Apply Theorem 5.25 with $f(x) = \frac{1}{x}$. Then

$$\int_a^b \frac{\varphi'(t)}{\varphi(t)}\,dt = \int_a^b f(\varphi(t))\varphi'(t)\,dt = \int_{\varphi(a)}^{\varphi(b)} \frac{1}{x}\,dx = \log(|x|)\Big|_{\varphi(a)}^{\varphi(b)} = \log(|\varphi(b)|) - \log(|\varphi(a)|).$$

□

Example 5.27. (1) We want to compute $\int_a^b \arctan x\,dx$. We first apply integration by parts with $u(x) = x$, $v(x) = \arctan x$, such that $u'(x) = 1$, $v'(x) = \frac{1}{1+x^2}$. Then

$$\int_a^b \arctan x\,dx = \int_a^b 1\cdot\arctan x\,dx = x\arctan x\Big|_a^b - \int_a^b \frac{x}{1+x^2}\,dx.$$

In the integral on the right-hand side, we substitute $\varphi(x) = 1 + x^2$, $\varphi'(x) = 2x$ and obtain:

$$\int_a^b \frac{x}{1+x^2}\,dx = \frac{1}{2}\int_a^b \frac{2x}{1+x^2}\,dx = \frac{1}{2}\int_{1+a^2}^{1+b^2} \frac{1}{y}\,dy. = \frac{1}{2}\log(1+b^2) - \frac{1}{2}\log(1+a^2).$$

Inserting this into the above, we obtain

$$\int_a^b \arctan x\,dx = \left(x\arctan x - \frac{1}{2}\log(1+x^2)\right)\Big|_a^b.$$

(2) We want to compute $\int_{-1}^1 \sqrt{1-x^2}\,dx$. One checks that the graph of $f : [-1,1] \to \mathbb{R}$, $f(x) = \sqrt{1-x^2}$ is the upper half circle of radius 1 around the origin, so from the geometric viewpoint, the result should be half of the area of a circle of radius 1 - which we know from high school to be $\frac{\pi}{2}$. Let's see.

This time we use Theorem 5.25 "the other way round". Let $\varphi : [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1,1]$, $\varphi(t) = \sin t$ and let $-1 \le a \le b \le 1$. Then, via integration by substitution, with $u := \arcsin a$ and $v := \arcsin b$ it holds that

$$\int_a^b \sqrt{1-x^2}\,dx = \int_u^v \sqrt{1-\sin^2(t)}\cos t\,dt = \int_u^v \sqrt{\cos^2(t)}\cos t\,dt = \int_u^v \cos^2(t)\,dt,$$

since $\cos t \ge 0$ for all $t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. By Theorem 3.78, $\cos^2(t) = \frac{1}{2}(\cos(2t)+1)$ holds for all $t \in \mathbb{R}$, hence

$$\begin{aligned}\int_a^b \sqrt{1-x^2}\,dx &= \frac{1}{2}\left(\int_u^v \cos(2t)\,dt + \int_{-\pi/2}^{\pi/2} 1\,dt\right)\\ &= \frac{1}{4}\int_{-2u}^{2v}\cos y\,dy + \frac{v-u}{2}\\ &= \frac{1}{4}\sin y\Big|_{2v}^{2u} + \frac{v-u}{2} = \left(\frac{1}{4}\sin(2\arcsin x) + \frac{1}{2}\arcsin x\right)\Big|_a^b.\end{aligned}$$

For each $t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ we compute by Theorem 3.78:

$$\sin(2t) = 2\sin t\cos t = 2\sin t\sqrt{1-\sin^2(t)}.$$

Consequently,

$$\sin(2\arcsin x) = 2x\sqrt{1-x^2} \qquad \forall x \in [-1,1].$$

Inserting this into the above computation yields

$$\int_a^b \sqrt{1-x^2}\,dx = \left(\frac{1}{2}x\sqrt{1-x^2} + \frac{1}{2}\arcsin x\right)\Big|_a^b.$$

In particular,

$$\int_{-1}^1 \sqrt{1-x^2}\,dx = \frac{1}{2}\sqrt{1-1} + \frac{1}{2}\arcsin(1) - \frac{1}{2}\sqrt{1-1} - \frac{1}{2}\arcsin(-1) = \frac{1}{2}\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = \frac{\pi}{2},$$

which coincides with our geometric reasoning. Yay!

5.4 Improper integrals

In this section, we want to study questions like the following one: Let $f : [a, +\infty) \to \mathbb{R}$ be a continuous function with $\lim_{x\to+\infty} f(x) = 0$. Geometrically, the graph of f approaches the x-axis as $x \to +\infty$. So if we interpret the integral as measuring the area below a graph, for very large values of r_1 and r_2, one might expect that the values of $\int_a^{r_1} f(x)\ dx$ and $\int_a^{r_2} f(x)\ dx$ are very close to each other. This leads to a simple question: if we view the integral $\int_a^r f(x)\ dx$ as a function in r, does this function converge as $r \to +\infty$? Questions like these are formalized by so-called improper integrals.

Definition 5.28. a) Let $a \in \mathbb{R}$, $b \in (a, +\infty) \cup \{+\infty\}$ and let $f : [a, b) \to \mathbb{R}$ with $f|_{[a,c]} \in \mathcal{R}([a, c])$ for every $c \in (a, b)$. If the following limit exists, then we put

$$\int_a^b f(x)\ dx := \lim_{r\nearrow b} \int_a^r f(x)\ dx$$

and say that $\int_a^b f(x)\ dx$ *converges*.

b) Let $b \in \mathbb{R}$, $a \in \{-\infty\} \cup (-\infty, b)$ and $f : (a, b] \to \mathbb{R}$, such that $f|_{[c,b]} \in \mathcal{R}([c, b])$ for every $c \in (a, b)$. If the following limit exists, then we put

$$\int_a^b f(x)\ dx := \lim_{r\searrow a} \int_r^b f(x)\ dx$$

and say that $\int_a^b f(x)\ dx$ *converges*.

c) If the limit does not exist in the situation of part a) or b), we say that $\int_a^b f(x)\ dx$ *diverges*.

d) Let $a \in \{-\infty\} \cup \mathbb{R}$, $b \in (a, +\infty) \cup \{+\infty\}$ and $f : (a, b) \to \mathbb{R}$, such that $f|_{[c,d]} \in \mathcal{R}([c, d])$ for all $c \in (a, b)$ and $d \in (c, b)$. Let $x_0 \in (a, b)$. If both $\int_a^{x_0} f(x)\ dx$ and $\int_{x_0}^b f(x)\ dx$ converge, then we put

$$\int_a^b f(x)\ dx := \int_a^{x_0} f(x)\ dx + \int_{x_0}^b f(x)\ dx$$

and say that $\int_a^b f(x)\ dx$ *converges*. (It follows from the computational rules for integrals that the value of $\int_a^b f(x)\ dx$ does not depend on the choice of x_0.)

Integrals of one of the forms introduced in a), b) and d) are called *improper integrals*.

Remark 5.29. Let $a, b \in \mathbb{R}$ with $a < b$. If $f : (a, b) \to \mathbb{R}$ is continuous and admits a continuous extension $\tilde{f} : [a, b] \to \mathbb{R}$, then it is derived from the fundamental theorem of calculus that

$$\int_a^b f(x)\ dx = \int_a^b \tilde{f}(x)\ dx.$$

Example 5.30. (1) Let $f : (-1, 1) \to \mathbb{R}$, $f(x) = \frac{1}{\sqrt{1-x^2}}$. One easily sees that

$$\lim_{x\to-1} f(x) = \lim_{x\to1} f(x) = +\infty.$$

However, we will show that $\int_{-1}^1 \frac{1}{\sqrt{1-x^2}}\ dx$ converges. We consider

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}}\ dx = \int_{-1}^0 \frac{1}{\sqrt{1-x^2}}\ dx + \int_0^1 \frac{1}{\sqrt{1-x^2}}\ dx$$

and compute, since arcsin is an antiderivative of f,

$$\int_0^1 \frac{1}{\sqrt{1-x^2}}\,dx = \lim_{r\nearrow 1}\int_0^r \frac{1}{\sqrt{1-x^2}}\,dx = \lim_{r\nearrow 1}\arcsin x\Big|_0^r = \lim_{r\nearrow 1}\arcsin r = \arcsin 1 = \frac{\pi}{2},$$
$$\int_{-1}^0 \frac{1}{\sqrt{1-x^2}}\,dx = \lim_{r\searrow -1}\int_r^0 \frac{1}{\sqrt{1-x^2}}\,dx$$
$$= \lim_{r\searrow -1}\arcsin x\Big|_r^0 = \lim_{r\searrow -1}(-\arcsin r) = -\arcsin(-1) = \frac{\pi}{2}.$$

Hence,

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}}\,dx = \frac{\pi}{2}+\frac{\pi}{2} = \pi.$$

Analogously, one may use the behaviour of the arctangent to derive that

$$\int_{-\infty}^{+\infty} \frac{1}{1+x^2}\,dx = \pi.$$

(2) Let $s \in \mathbb{R}\setminus\{-1\}$ and consider the improper integral $\int_1^{+\infty}\frac{1}{x^s}\,dx$. We compute for $r>1$ that

$$\int_1^r \frac{1}{x^s}\,dx = \frac{1}{1-s}x^{1-s}\Big|_1^r = \frac{1}{1-s}\left(\frac{1}{r^{s-1}}-1\right) = \frac{1}{s-1}\left(1-\frac{1}{r^{s-1}}\right).$$

Hence,

$$\int_1^{+\infty}\frac{1}{x^s}\,dx = \lim_{r\to+\infty}\int_1^r\frac{1}{x^s}\,dx = \lim_{r\to+\infty}\frac{1}{s-1}\left(1-\frac{1}{r^{s-1}}\right).$$

Since $\lim_{r\to+\infty}\frac{1}{r^{s-1}}$ exists if and only if $s-1>0$, one derives that $\int_1^{+\infty}\frac{1}{x^s}\,dx$ converges if and only if $s>1$. In this case,

$$\int_1^{+\infty}\frac{1}{x^s}\,dx = \frac{1}{s-1}.$$

(3) Let again $s\in\mathbb{R}\setminus\{-1\}$ and consider $\int_0^1\frac{1}{x^s}\,dx$. For $r\in(0,1)$ we compute

$$\int_r^1\frac{1}{x^s}\,dx = \frac{x^{1-s}}{1-s}\Big|_r^1 = \frac{1}{1-s}(1-r^{1-s}).$$

Since $\lim_{r\to 0}r^{1-s}$ exists if and only if $1-s>0$, we derive that $\int_0^1\frac{1}{x^s}\,dx$ converges if and only if $s<1$. In that case,

$$\int_0^1\frac{1}{x^s}\,dx = \lim_{r\searrow 0}\int_r^1\frac{1}{x^s}\,dx = \lim_{r\searrow 0}\frac{1}{1-s}(1-r^{1-s}). = \frac{1}{1-s}.$$

(4) Using the properties of $\log x$ one further sees that both $\int_0^1\frac{1}{x}\,dx$ and $\int_1^{+\infty}\frac{1}{x}\,dx$ diverge.

Theorem 5.31. *Let $a\in\mathbb{R}$, $b\in(a,+\infty)\cup\{+\infty\}$ and let $f:[a,b)\to\mathbb{R}$ with $f|_{[a,c]}\in\mathcal{R}([a,c])$ for all $c\in(a,b)$. Assume that there is a function $\varphi:[a,b)\to[0,+\infty)$ with $|f(x)|\le\varphi(x)$ for all $x\in[a,b)$ and for which $\int_a^b\varphi(x)\,dx$ converges. Then $\int_a^b f(x)\,dx$ converges.*

Proof. Let $F, \Phi : [a,b) \to \mathbb{R}$, $F(r) := \int_a^r f(x)\ dx$, $\Phi(r) := \int_a^r \varphi(x)\ dx$. We need to show that $\lim_{t \nearrow b} F(t)$ exists. For all $s, t \in (a,b)$ we compute using Theorems 5.11.c) and 5.13.b) that

$$|F(t) - F(s)| = \left| \int_s^t f(x) \right| \leq \int_s^t |f(x)|\ dx \leq \left| \int_s^t \varphi(x)\ dx \right| = |\Phi(t) - \Phi(s)|\,. \tag{5.5}$$

Let $(t_n)_{n\in\mathbb{N}}$ be a sequence in $[a,b)$ with $\lim_{n\to\infty} t_n = b$. By assumption, $(\Phi(t_n))_{n\in\mathbb{N}}$ converges, thus is a Cauchy sequence. By (5.5), $(F(t_n))_{n\in\mathbb{N}}$ is a Cauchy sequence as well and thus convergent in $\mathbb{R}$. It only remains to be shown that $\lim_{n\to\infty} F(t_n)$ is independent of the choice of $(t_n)_{n\in\mathbb{N}}$. If $(s_n)_{n\in\mathbb{N}}$ is another sequence in $[a,b)$ with $\lim_{n\to\infty} s_n = b$, then

$$\lim_{n\to\infty} |F(t_n) - F(s_n)| \overset{(5.5)}{\leq} \lim_{n\to\infty} |\Phi(t_n) - \Phi(s_n)| = 0$$

by convergence of Φ. Thus, $\lim_{n\to\infty} F(s_n) = \lim_{n\to\infty} F(t_n)$, which shows the claim. □

In the next result, we revisit the convergence of series. It turns out that using improper integrals we can establish another convergence test for series.

Theorem 5.32 (Integral test)**.** *Let $f : [1,+\infty) \to [0,+\infty)$ be monotonically decreasing. Then $\sum_{n=1}^{\infty} f(n)$ converges if and only if $\int_1^{+\infty} f(x)\ dx$ converges.*

Proof. Define $g_1, g_2 : [1,+\infty) \to [0,+\infty)$ by

$$g_1(x) = f(n+1), \quad g_2(x) = f(n) \quad \forall x \in [n, n+1),\ \ n \in \mathbb{N}.$$

Since f is monotonically decreasing, it holds that $f(n+1) \leq f(x) \leq f(n)$ for all $n \in \mathbb{N}$ and $x \in [n, n+1)$. This yields

$$g_1(x) \leq f(x) \leq g_2(x) \qquad \forall x \in [1,+\infty).$$

Moreover, in analogy with Example 5.6, one shows that for each $N \in \mathbb{N}$ with $N \geq 2$, it holds that

$$\int_1^N g_1(x)\ dx = \sum_{n=1}^{N-1} f(n+1) = \sum_{n=2}^{N} f(n), \qquad \int_1^N g_2(x)\ dx = \sum_{n=1}^{N-1} f(n).$$

and thus by Theorem 5.11.c)

$$\sum_{n=2}^{N} f(n) \leq \int_1^N f(x)\ dx \leq \sum_{n=1}^{N-1} f(n) \qquad \forall N \in \mathbb{N},\ N > 1. \tag{5.6}$$

If $\int_1^{\infty} f(x)\ dx$ converges, this yields that $\sum_{n=2}^{N} f(n) \leq \int_1^{+\infty} f(x)\ dx$ for all $N \geq 2$, hence $\sum_{n=1}^{\infty} f(n)$ converges by Theorem 2.66.
Conversely, if $\sum_{n=1}^{\infty} f(n)$ converges, it follows that $(\int_1^N f(x)\ dx)_{N\in\mathbb{N}}$ is bounded from above. Since f is non-negative, the sequence is monotonically increasing as well, hence convergent.
For $x \in \mathbb{R}$ let again $\lfloor x \rfloor = \sup\{k \in \mathbb{Z} \mid n \leq x\}$ and put $\lceil x \rceil := \inf\{k \in \mathbb{Z} \mid x \leq k\}$. Letf $(r_n)_{n\in\mathbb{N}}$ be an arbitrary sequence in $[1,+\infty)$ with $\lim_{n\to\infty} r_n = +\infty$. Then $(\int_1^{\lfloor r_n \rfloor} f(x)\ dx)_{n\in\mathbb{N}}$ and $(\int_1^{\lceil r_n \rceil} f(x)\ dx)_{n\in\mathbb{N}}$ are subsequences of $(\int_1^N f(x)\ dx)_{N\in\mathbb{N}}$, thus they are convergent with

$$\lim_{n\to\infty} \int_1^{\lfloor r_n \rfloor} f(x)\ dx = \lim_{n\to\infty} \int_1^{\lceil r_n \rceil} f(x)\ dx = \lim_{N\to\infty} \int_1^N f(x)\ dx.$$

Since f is non-negative,

$$\int_1^{\lfloor r_n \rfloor} f(x)\,dx \leq \int_1^{r_n} f(x)\,dx \leq \int_1^{\lceil r_n \rceil} f(x)\,dx \qquad \forall n \in \mathbb{N},$$

so it follows from the squeeze theorem that $\lim_{n\to\infty} \int_1^{r_n} f(x)\,dx = \lim_{N\to\infty} \int_1^N f(x)\,dx$. Since $(r_n)_{n\to\mathbb{N}}$ was chosen arbitrarily, this shows that $\int_1^{+\infty} f(x)\,dx$ converges. □

Example 5.33. (1) Combining our observations from Example 5.30 with Theorem 5.32, we can generalize an observation from Chapter 3 and show:

$$\sum_{n=1}^{\infty} \frac{1}{n^s} \text{ converges} \quad \Leftrightarrow \quad \int_1^{+\infty} \frac{1}{x^s}\,dx \text{ converges} \quad \Leftrightarrow \quad s > 1.$$

(2) Consider the series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{(n+1)(\log(n+1))^s}$, where $s \in \mathbb{R}$. Since $f : [1, +\infty) \to [0, +\infty)$, $f(x) = \frac{1}{(x+1)(\log(x+1))^s}$ is monotonically decreasing, the integral test implies that the series converges if and only if

$$\int_1^{+\infty} \frac{1}{(x+1)(\log(x+1))^s}\,dx$$

converges. For $r > 1$ we compute by substitution of $\varphi(x) = \log(x+1)$, $\varphi'(x) = \frac{1}{x+1}$ that

$$\int_1^r \frac{1}{(x+1)(\log(x+1))^s}\,dx = \int_{\log 2}^{\log(r+1)} \frac{1}{u^s}\,ds.$$

One derives that $\int_1^{+\infty} \frac{1}{(x+1)(\log(x+1))^s}\,dx = \int_{\log 2}^{+\infty} \frac{1}{u^s}\,du$, so it follows from Example 5.30.(1) that the integral converges if and only if $s > 1$. Hence,

$$\sum_{n=1}^{\infty} \frac{1}{(n+1)(\log(n+1))^s} \text{ converges} \quad \Leftrightarrow \quad s > 1.$$

(3) We want to study the methods of proof of Theorem 5.32 for $f(x) = \frac{1}{x}$ in greater detail. Since $\int_1^N \frac{1}{x}\,dx = \log N - \log 1 = \log N$, the inequalities (5.6) then read as

$$\sum_{n=2}^{N} \frac{1}{n} \leq \log N \leq \sum_{n=1}^{N-1} \frac{1}{n} \qquad \forall n \in \mathbb{N}.$$

Roughly, this says that the sequence of partial sums $(\sum_{n=1}^N \frac{1}{n})_{N\in\mathbb{N}}$ of the harmonic series diverges to $+\infty$ as fast (or as slow) as $\log N$ as $N \to +\infty$. For a more precise statement, put

$$\gamma_N := \sum_{n=1}^{N} \frac{1}{n} - \log N \qquad \forall N \in \mathbb{N}.$$

The above inequalities imply $-\sum_{n=1}^{N-1} \frac{1}{n} \leq -\log N \leq -\sum_{n=2}^{N} \frac{1}{n}$ for all $N \in \mathbb{N}$. Adding $\sum_{n=1}^N \frac{1}{n}$ yields

$$\frac{1}{N} \leq \gamma_N \leq 1 \qquad \forall N \in \mathbb{N},$$

so $(\gamma_N)_{N\in\mathbb{N}}$ is bounded. Moreover, for each $N \in \mathbb{N}$ we compute

$$\begin{aligned}\gamma_{N+1} - \gamma_N &= \frac{1}{N+1} - \log(N+1) + \log N \\ &= \frac{1}{N+1} - \int_N^{N+1} \frac{1}{x}\, dx = \int_N^{N+1} \left(\frac{1}{N+1} - \frac{1}{x}\right)\, dx \leq 0,\end{aligned}$$

since $\frac{1}{x} \geq \frac{1}{N+1}$ for all $x \in [N, N+1]$. Thus, $(\gamma_N)_{N\in\mathbb{N}}$ is monotonically decreasing and bounded, hence convergent. Its limit

$$\gamma := \lim_{N\to\infty} \left(\sum_{n=1}^{N} \frac{1}{n} - \log N\right)$$

is called the *Euler-Mascheroni constant*. To this day, it is an open problem in mathematics to find out whether γ is a rational or an irrational number.[1]

Dear reader,

you have reached the part of the lecture notes that will not be relevant for the written exam to this course. Everything above the double line might be part of the exam, while everything below will definitely not be part of the exam. You are however encouraged to continue reading, since there are beautiful computations waiting for you in the following section.

5.5 Integration of rational functions

There is an old saying among mathematicians that

"Differentiation is mechanics, integration is art."

Knowing the derivatives of the most elementary functions, the product rule and the chain rule, one is able to derive all functions that are given in terms of these elementary function by simply following the rules of differentiation.
Integrating an arbitrary function tends to be way more complicated and the knowledge of integrals of elementary functions does not at all enable us to integrate arbitrary functions that are given in terms of them. There are even functions, for example e^{-x^2}, for which there is no explicit form of their antiderivatives. It is also not straightforward to find the right approach of which method of integration to us and it might be very difficult to see which trick to use[2].

In this section, we will consider integrals of rational functions, which is an important class of functions that so far we have not studied in detail.

[1] As a number, γ is approximately given by $\gamma = 0.5772156649...$ In May 2020, the first 600,000,000,100 digits of the value of γ have been computed by (bored) mathematicians.

[2] See also `https://xkcd.com/2117/`.

Definition 5.34. A (real) *rational function* is a function of the form

$$f : D \to \mathbb{R}, \qquad f(x) = \frac{p(x)}{q(x)},$$

where $p, q : \mathbb{R} \to \mathbb{R}$ are polynomials and where $D \subset \mathbb{R} \setminus \{x \in \mathbb{R} \mid q(x) = 0\}$.

By our rules of differentiation, we know that any such rational function is differentiable with

$$f'(x) = \frac{p'(x)q(x) - p(x)q'(x)}{(q(x))^2} \quad \forall x \in D,$$

which is again a rational function.

In contrast, an antiderivative of a rational function is much harder to find and not necessarily a rational function as the simple examples of $f(x) = \frac{1}{x}$ and $f(x) = \frac{1}{1+x^2}$ already show. In the following, we will carry out step by step how to find integrals of an arbitrary rational function $f(x) = \frac{p(x)}{q(x)}$. We will further denote the degree of a polynomial P by $\deg P$.

Step 1: degree reduction

First of all we observe that if $\deg q = 0$, i.e. q is constant, then f is a polynomial, which we can already integrate. It remains to consider the case of $\deg q > 0$.

Theorem 5.35. *Assume that* $\deg p \geq \deg q > 0$. *Then there are polynomials* g *and* p_1 *with* $\deg p_1 < \deg q$, *such that*

$$p(x) = g(x) \cdot q(x) + p_1(x) \qquad \forall x \in \mathbb{R}.$$

This theorem can be shown using elementary methods, but here we refrain from giving a proof. With r and p_1 as in the theorem, we obtain for $[a, b] \subset D$ that

$$\int_a^b f(x)\, dx = \int_a^b \frac{p(x)}{q(x)}\, dx = \int_a^b g(x)\, dx + \int_a^b \frac{p_1(x)}{q(x)}\, dx.$$

Since g is easily integrated as a polynomial, it suffices to find a way to integrate $\int_a^b \frac{p_1(x)}{q(x)}\, dx$. The difference to the original setting is now that the degree of the numerator polynomial is smaller than the one of the denominator polynomial.
Thus, in the following we will only consider the case of $f(x) = \frac{p(x)}{q(x)}$ with $\deg p < \deg q$.

Step 2: summary of known results

We have seen various examples of integrals of rational functions $f(x) = \frac{p(x)}{q(x)}$ in the previous sections. If q is constant, then f is a polynomial which is easily integrated. If p is constant, then we can use our previous knowledge to integrate the function in some cases.
Combining our previous computations for $\int_a^b \frac{1}{x^k}\, dx$ and $\int_a^b \frac{1}{1+x^2}\, dx$ with Corollary 5.26.a) we derive for arbitrary $c \in \mathbb{R}$ that

- $\displaystyle\int_a^b \frac{1}{(x-c)^k}\, dx = \frac{1}{(1-k)(x-c)^{k-1}}\Big|_a^b \qquad$ if $k > 1$, $c \notin [a, b]$.

- $\displaystyle\int_a^b \frac{1}{x-c}\,dx = \log(|x-c|)\Big|_a^b \qquad \text{if } c \notin [a,b].$
- $\displaystyle\int_a^b \frac{1}{1+(x-c)^2}\,dx = \arctan(x-c)\Big|_a^b.$

These computations will be the building blocks for our discussion of the general case.

Step 3: rational functions with quadratic denominator

In the case of $\deg q = 1$, p is constant and this case is covered by the integrals from Step 2. In this step we take a closer look at the case of $\deg q = 2$, such that the desired integral has the general form

$$\int_a^b \frac{\mu x + \nu}{x^2 + 2rx + s}\,dx,$$

where $\mu, \nu, r, s \in \mathbb{R}$. We further assume in this step *that $x^2 + 2rx + s$ has no solutions in $\mathbb{R}$, or equivalently, that $s > r^2$*. (The general case will be considered in Step 4.)

With $q(x) = x^2 + 2rx + s$ we compute that $q'(x) = 2x + 2r$. Assume that $\mu \neq 0$. Then

$$\mu x + \nu = \frac{\mu}{2}\left(2x + \frac{2\nu}{\mu}\right) = \frac{\mu}{2}\left(q'(x) + \frac{2\nu}{\mu} - 2r\right).$$

Hence, using Corollary 5.26.d),

$$\begin{aligned}
\int_a^b \frac{\mu x + \nu}{x^2 + 2rx + s}\,dx &= \frac{\mu}{2}\int_a^b \frac{q'(x)}{q(x)}\,dx + \frac{\mu}{2}\int_a^b \frac{\frac{2\nu}{\mu} - 2r}{x^2 + 2rx + s}\,dx \\
&= \frac{\mu}{2}\log(x^2 + 2rx + s)\Big|_a^b + \int_a^b \frac{\nu - r\mu}{x^2 + 2rx + s}\,dx \\
&= \frac{\mu}{2}\log\left(\frac{b^2 + 2rb + s}{a^2 + 2ra + s}\right) + (\nu - r\mu)\int_a^b \frac{1}{x^2 + 2rx + s}\,dx.
\end{aligned}$$

One immediately sees that this last formula is valid in the case of $\mu = 0$ as well. To compute the remaining integral, we rewrite $x^2 + 2rx + s = (x+r)^2 + s - r^2$. Then

$$\begin{aligned}
\int_a^b \frac{1}{x^2 + 2rx + s}\,dx &= \int_a^b \frac{1}{s - r^2 + (x+r)^2}\,dx \\
&= \frac{1}{s-r^2}\int_a^b \frac{1}{1 + \frac{(x+r)^2}{s-r^2}}\,dx = \frac{1}{s-r^2}\int_a^b \frac{1}{1 + \left(\frac{x+r}{\sqrt{s-r^2}}\right)^2}\,dx.
\end{aligned}$$

Now substitute $\varphi(x) = \frac{x+r}{\sqrt{s-r^2}}$, $\varphi'(x) = \frac{1}{\sqrt{s-r^2}}$. Then

$$\begin{aligned}
\int_a^b \frac{1}{x^2 + 2rx + s}\,dx &= \frac{1}{\sqrt{s-r^2}}\int_{(a+r)/(\sqrt{s-r^2})}^{(b+r)/(\sqrt{s-r^2})} \frac{1}{1+u^2}\,du \\
&= \frac{1}{\sqrt{s-r^2}}\left(\arctan\left(\frac{b+r}{\sqrt{s-r^2}}\right) - \arctan\left(\frac{a+r}{\sqrt{s-r^2}}\right)\right) \\
&= \frac{1}{\sqrt{s-r^2}}\arctan\left(\frac{b+r}{\sqrt{s-r^2}}\right)\Big|_a^b.
\end{aligned}$$

Inserting this into the above computation yields:

$$\int_a^b \frac{\mu x+\nu}{x^2+2rx+s}\,dx = \frac{\mu}{2}\log(x^2+2rx+s)\Big|_a^b + \frac{\nu-r\mu}{\sqrt{s-r^2}}\arctan\left(\frac{x+r}{\sqrt{s-r^2}}\right)\Big|_a^b. \tag{5.7}$$

Not exactly obvious and not exactly beautiful to look at, but a very interesting computation (in the lecturer's point of view).

Step 4: partial fraction decomposition

The general case of a rational function $f(x) = \frac{p(x)}{q(x)}$ with $\deg p < \deg q$ can now be derived by combining the results of Steps 2 and 3 with an elementary decomposition result for rational functions. We begin by presenting an elementary factorization result for polynomials whose proof we omit.

Theorem 5.36 (Factor decomposition of real polynomials)**.** *Let $q : \mathbb{R} \to \mathbb{R}$ be a real polynomial of degree $n \in \mathbb{N}$. Then there exist $k \in \{1, 2, \dots, n\}$ and real polynomials $q_1, \dots, q_k$ with $\deg q_j \leq 2$ for all $j \in \{1, 2, \dots, k\}$, such that*

$$q(x) = q_1(x) \cdot q_2(x) \cdot \ldots \cdot q_k(x) \qquad \forall x \in \mathbb{R}$$

and such that $q_j(x) \neq 0$ for all $x \in \mathbb{R}$ if $\deg q_j = 2$.

Remark 5.37. One shows that if $a \in \mathbb{R}$ satisfies $q(a) = 0$, then q_1 in Theorem 5.36 can be chosen as $q_1(x) = x - a$. By polynomial long divison, a technique you might be familiar with from high school, one can find an explicit polynomial p_1 with $p = q_1 \cdot p_1$. One may then iterate the process by finding $a_1 \in \mathbb{R}$ with $p_1(a_1) = 0$.

Example 5.38. (1) Let $p(x) = x^4 - 1$. Then

$$p(x) = (x^2+1)(x^2-1) = (x^2+1)(x+1)(x-1) \qquad x \in \mathbb{R}.$$

which has the properties of Theorem 5.36, since $x^2 + 1 \neq 0$ for all $x \in \mathbb{R}$.

(2) Let $p(x) = x^3 - 1$. Then $p(1) = 0$. The geometric sum formula yields

$$\frac{x^3-1}{x-1} = \frac{1-x^3}{1-x} = x^2 + x + 1,$$

hence $x^3 - 1 = (x-1)(x^2+x+1)$ for all $x \in \mathbb{R}$. Since

$$x^2 + x + 1 = 0 \quad \Leftrightarrow \quad (x+\tfrac{1}{2})^2 + \tfrac{3}{4} = 0,$$

the equation $x^2 + x + 1 = 0$ has no real solutions, so our decomposition satisfies the conditions of Theorem 5.36.

For our rational function $f(x) = \frac{p(x)}{q(x)}$ we can apply the factor decomposition theorem to q, such that f is of the form

$$f(x) = \frac{p(x)}{q_1(x) \cdot q_2(x) \cdot \ldots \cdot q_k(x)},$$

where $\deg q_j \leq 2$ and $q_j(x) \neq 0$ for all $x \in \mathbb{R}$ if $\deg q_j = 2$ for all $j \in \{1, 2, \ldots, k\}$. Explicitly, this means that there are $m, r \in \mathbb{N}$, $k_1, \ldots, k_m \in \mathbb{N}$ and $c, a_1, \ldots, a_m, b_1, \ldots, b_r, c_1, \ldots, c_r \in \mathbb{R}$, such that

$$q(x) = c \cdot (x - a_1)^{k_1}(x - a_2)^{k_2} \ldots (x - a_m)^{k_m} \cdot (x^2 + 2b_1x + c_1) \cdot \ldots \cdot (x^2 + 2b_rx + c_r).$$

The following theorem now provides a way of using this decomposition to write f as a sum of rational functions with much simpler denominators.

Theorem 5.39 (Partial fraction decomposition)**.** *In the above situation, there are numbers $\lambda_{i,j} \in \mathbb{R}$, $i \in \{1, 2, \ldots, m\}$, $j \in \{1, 2, \ldots, k_i\}$, $\mu_1, \ldots, \mu_r, \nu_1, \ldots, \nu_r \in \mathbb{R}$, such that*

$$f(x) = \underbrace{c \sum_{j=1}^{k_1} \frac{\lambda_{1,j}}{(x - a_1)^j} + \cdots + c \sum_{j=1}^{k_m} \frac{\lambda_{m,j}}{(x - a_m)^j}}_{=:f_1(x)} + \underbrace{c \sum_{i=1}^{r} \frac{\mu_i x + \nu_i}{x^2 + 2b_i x + c_i}}_{=:f_2(x)}.$$

The part of f denoted by f_1 might look complicated, but this part can now directly be integrated, since all of its summands are of a form that we considered in Step 2:

$$\begin{aligned}\int_a^b f_1(x)\, dx &= \sum_{i=1}^{m} \int_a^b \left(\sum_{j=1}^{k_i} \frac{\lambda_{i,j}}{(x - a_i)^j} \right) dx = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \lambda_{i,j} \int_a^b \frac{1}{(x - a_i)^j}\, dx \\ &\overset{\text{Step 2}}{=} \sum_{i=1}^{m} \lambda_{i,1} \log(|x - a_i|) \Big|_a^b + \sum_{i=1}^{m} \sum_{j=2}^{k_i} \frac{\lambda_{i,j}}{(1 - j)(x - a_i)^{j-1}} \Big|_a^b.\end{aligned}$$

It remains to integrate the summands of the function f_2 from Theorem 5.39. All summands of f_2 are of a form considered in Step 3. Using (5.7), we obtain

$$\begin{aligned}\int_a^b f_2(x)\, dx &= \sum_{i=1}^{r} \int_a^b \frac{\mu_i x + \nu_i}{x^2 + 2b_i x + c}\, dx \\ &= \sum_{i=1}^{r} \left(\frac{\mu_i}{2} \log(x^2 + 2b_i x + c_i) \Big|_a^b + \frac{\nu_i - b_i \mu_i}{\sqrt{c_i - b_i^2}} \arctan\left(\frac{x + b_i}{\sqrt{c_i - b_i^2}} \right) \Big|_a^b \right).\end{aligned}$$

Combining the two computations for the integrals of f_1 and f_2, one finally obtains an explicit formula for $\int_a^b f(x)\, dx$, so following Steps 1 to 4, we can compute the integral of any rational function.

Example 5.40. (1) Let

$$f : \mathbb{R} \setminus \{-1, 1\} \to \mathbb{R}, \qquad f(x) = \frac{1}{x^4 - 1},$$

and let $a < b \in \mathbb{R}$ with $[a, b] \subset \mathbb{R} \setminus \{-1, 1\}$. Using the computation of Example 5.38.(1), we want to compute its partial fraction decomposition which is of the form

$$f(x) = \frac{A}{x^2 + 1} + \frac{B}{x + 1} + \frac{C}{x - 1}.$$

We compute explicitly that

$$\begin{aligned} f(x) = \frac{1}{(x^2+1)(x^2-1)} &= \frac{1}{2}\left(\frac{1}{x^2-1} - \frac{1}{x^2+1}\right) \\ &= \frac{1}{2}\left(\frac{1}{2}\left(\frac{1}{x-1} - \frac{1}{x+1}\right) - \frac{1}{x^2+1}\right) \\ &= \frac{1}{4(x-1)} - \frac{1}{4(x+1)} - \frac{1}{2(x^2+1)}, \end{aligned}$$

and thus

$$\begin{aligned} \int_a^b \frac{1}{x^4-1}\,dx &= \frac{1}{4}\int_a^b \frac{1}{x-1}\,dx - \frac{1}{4}\int_a^b \frac{1}{x+1}\,dx - \frac{1}{2}\int_a^b \frac{1}{1+x^2}\,dx \\ &= \left(\frac{1}{4}\log|x-1| - \frac{1}{4}\log|x+1| - \frac{1}{2}\arctan x\right)\Big|_a^b. \end{aligned}$$

(2) Let $f : \mathbb{R} \setminus \{1\} \to \mathbb{R}$, $f(x) = \frac{1}{x^3-1}$ and let $a < b \in \mathbb{R}$ with $[a,b] \subset \mathbb{R} \setminus \{1\}$. By Example 5.38.(2), the partial fraction decomposition of f is of the form

$$f(x) = \frac{A}{x-1} + \frac{Bx+C}{x^2+x+1}$$

for some $A, B, C \in \mathbb{R}$. By multiplying the equation with $x^3 - 1$ we obtain that

$$1 = A(x^2+x+1) + (Bx+C)(x-1) = (A+B)x^2 + (A-B+C)x + A - C \qquad \forall x \in \mathbb{R},$$

which implies

$$A + B = 0 \quad \wedge \quad A - B + C = 0 \quad \wedge A - C = 1.$$

By a technique you will learn in Mathematics 2, one derives that $A = \frac{1}{3}$, $B = -\frac{1}{3}$,$C = -\frac{2}{3}$, such that

$$f(x) = \frac{1}{3(x-1)} - \frac{x+2}{3(x^2+x+1)} \qquad \forall x \in \mathbb{R} \setminus \{1\}.$$

Hence,

$$\begin{aligned} \int_a^b \frac{1}{x^3-1}\,dx &= \frac{1}{3}\int_a^b \frac{1}{x-1}\,dx - \frac{1}{3}\int_a^b \frac{x+2}{x^2+x+1}\,dx \\ &= \frac{1}{3}\log(|x-1|)\Big|_a^b - \frac{1}{3}\int_a^b \frac{x+2}{x^2+x+1}\,dx. \end{aligned}$$

Applying (5.7) with $\mu = 1$, $\nu = 2$, $r = \frac{1}{2}$ and $s = 1$, we compute that

$$\begin{aligned} \int_a^b \frac{x+2}{x^2+x+1}\,dx &= \frac{1}{2}\log(x^2+x+1)\Big|_a^b + \frac{2-\frac{1}{2}}{\sqrt{1-\frac{1}{4}}}\arctan\left(\frac{x+\frac{1}{2}}{\sqrt{1-\frac{1}{4}}}\right)\Big|_a^b \\ &= \frac{1}{2}\log(x^2+x+1)\Big|_a^b + \sqrt{3}\arctan\left(\frac{2x+1}{\sqrt{3}}\right)\Big|_a^b. \end{aligned}$$

Combining the previous computations, we obtain

$$\int_a^b \frac{1}{x^3-1}\,dx = \left(\frac{1}{3}\log(|x-1|) - \frac{1}{6}\log(x^2+x+1) - \frac{1}{\sqrt{3}}\arctan\left(\frac{2x+1}{\sqrt{3}}\right)\right)\Big|_a^b.$$

Depending on one's personal taste, one might reformulate the logarithms on the right-hand side in a more conside way. By the computational rules for logarithms,

$$\begin{aligned}\frac{1}{3}\log(|x-1|)-\frac{1}{6}\log(x^2+x+1) &= \frac{1}{3}\left(\log(|x-1|)-\log(\sqrt{x^2+x+1})\right)\\ &= \frac{1}{3}\log\left(\frac{|x-1|}{\sqrt{x^2+x+1}}\right),\end{aligned}$$

such that

$$\int_a^b \frac{1}{x^3-1}\,dx = \left(\frac{1}{3}\log\left(\frac{|x-1|}{\sqrt{x^2+x+1}}\right)-\frac{1}{\sqrt{3}}\arctan\left(\frac{2x+1}{\sqrt{3}}\right)\right)\Bigg|_a^b.$$

These examples hopefully underline the saying from the beginning: "Differentiation is mechanics, integration is art."

Chapter 6

Vector spaces

6.1 Vector spaces and linear subspaces

This chapter is the first of the mathematics course that explicitly deals with linear algebra. In your theoretical physics course, you have studies points in $\mathbb{R}^n$ as vectors and discussed the addition of vectors, their multiplications by scalars and the geometric intuition behind them. We will now introduce an abstract setting of so-called vector spaces that generalizes the corresponding properties of $\mathbb{R}^n$. While it might not be apparent for you at this point why physicists are interested in studying other vector spaces than $\mathbb{R}^n$ or $\mathbb{C}^n$, it will become clear to you in future lectures that e.g. in quantum mechanics certain infinite-dimensional vector spaces will be required to properly formulate some results mathematically.

Definition 6.1. Let K be a field. An tuple $(V, +, m)$ is called a *vector space over K* (or a *K-vector space*) if $(V, +)$ is an abelian group and if $m : K \times V \to V$ is a map, usually denoted by $\lambda \cdot v := m(\lambda, v)$ or simply by λv, for all $\lambda \in K$ and $v \in V$, with the following properties:

(i) $(\lambda + \mu) \cdot v = \lambda \cdot v + \mu \cdot v$ for all $\lambda, \mu \in K$ and $v \in V$.

(ii) $\lambda \cdot (v + w) = \lambda \cdot v + \lambda \cdot w$ for all $\lambda \in K$ and $v, w \in V$.

(iii) $\lambda \cdot (\mu \cdot v) = (\lambda\mu) \cdot v$ for all $\lambda, \mu \in K$ and $v \in V$.

(iv) $1 \cdot v = v$ for all $v \in V$, where $1 \in K$ is the neutral element of multiplication.

m is called the *scalar multiplication of V*. Elements of a vector space are also called *vectors*. A vector space over $\mathbb{R}$ is called a *real vector space* and a vector space over $\mathbb{C}$ is called a *complex vector space*. One mostly just writes V instead of $(V, +, m)$ and assumes that addition and scalar multiplication are defined on V.

Remark 6.2. In most applications in physics, one only studies the cases of $K = \mathbb{R}$ and $K = \mathbb{C}$. Vector spaces over other fields are nevertheless useful in several other applications, e.g. in cryptography, where parts of it can be formulated in terms of vector spaces over fields with finitely many elements.

Example 6.3. Let K be a field.

(1) For each $n \in \mathbb{N}$

$$K^n = \underbrace{K \times \cdots \times K}_{n \text{ times}} = \{(x_1, \ldots, x_n) \mid x_1, \ldots, x_n \in K\}$$

is a vector space over K with the operations

$$(x_1, x_2, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n),$$
$$\lambda \cdot (x_1, x_2, \ldots, x_n) := (\lambda x_1, \lambda x_2, \ldots, \lambda x_n)$$

for all $\lambda, x_1, \ldots, x_n, y_1, \ldots, y_n \in K$. One derives from the field axioms that $(V, +)$ is an abelian group and that $\cdot$ is indeed a scalar multiplication. In particular, $K = K^1$ is a vector space over itself.

(2) Let V be a vector space over K and let M be a set. Then

$$V^M := \{f \mid f : M \to V \text{ is a map}\}$$

is a vector space over K via

$$(f + g)(x) := f(x) + g(x) \qquad (\lambda \cdot f)(x) := \lambda \cdot f(x)$$

for all $f, g \in V^M$, $\lambda \in K$ and $x \in M$. The vector space properties of V^M can hereby be derived from the vector space properties of V.

(3) As a special case of (2), consider $\mathbb{R}^{\mathbb{N}}$ as a real vector space. By definition,

$$\mathbb{R}^{\mathbb{N}} = \{(a_n)_{n \in \mathbb{N}} \mid a_n \in \mathbb{R}\ \forall n \in \mathbb{N}\},$$

i.e. $\mathbb{R}^{\mathbb{N}}$ is the vector space of real sequences.

Similar as we did with fields in Section 1.2, we can derive further computational properties of vector spaces straight from the axioms.

Theorem 6.4. *Let K be a field, let V be a vector space over K and let $0_K \in K$ and $0_V \in V$ denote the respective neutral elements of addition. Then:*

a) $0_K \cdot v = 0_V$ *for all* $v \in V$.

b) $\lambda \cdot 0_V = 0_V$ *for all* $\lambda \in K$.

c) $(-1) \cdot v = -v$ *for all* $v \in V$.

d) Let $\lambda \in K$ *and* $v \in V$ *with* $\lambda \cdot v = 0_V$. *Then* $\lambda = 0_K$ *or* $v = 0_V$.

Proof. Let $\lambda \in K$ and $v \in V$. Then (where small roman numerals refer to Definition 6.1):

a) $0_K \cdot v = (0_K + 0_K) \cdot v \overset{\text{(i)}}{=} 0_K \cdot v + 0_K \cdot v$, which shows the claim.[1]

b) $\lambda \cdot 0_V = \lambda \cdot (0_V + 0_V) \overset{\text{(ii)}}{=} \lambda \cdot 0_V + \lambda \cdot 0_V$, which shows the claim.

c) $v + (-1) \cdot v \overset{\text{(iv)}}{=} 1 \cdot v + (-1) \cdot v \overset{\text{(i)}}{=} (1 + (-1)) \cdot v = 0_K \cdot v \overset{\text{a)}}{=} 0_V$. Analogously, one shows that $(-1) \cdot v + v = 0$, such that $(-1) \cdot v = -v$.

d) Assume that $\lambda \cdot v = 0_V$, but $\lambda \neq 0_K$. Then

$$v \overset{\text{(iv)}}{=} 1 \cdot v = (\lambda^{-1}\lambda) \cdot v \overset{\text{(iii)}}{=} \lambda^{-1} \cdot (\lambda \cdot v) = \lambda^{-1} \cdot 0_V \overset{\text{b)}}{=} 0_V.$$

[1] If this argument looks familiar to you, compare it with the proof of Theorem 1.32.

□

In the following we will denote both 0_K and 0_V simply by 0 in case it is clear from the context which of the two we are referring to.

Definition 6.5. Let K be a field and V be a vector space over K. A non-empty subset $U \subset V$ is called a *linear subspace of V* (or *vector subspace of V*) if it satisfies the following conditions:

(i) For all $v, w \in U$ it holds that $v + w \in U$.

(ii) For all $\lambda \in K$ and $v \in U$ it holds that $\lambda v \in U$.

Remark 6.6. By definition of linear subspaces, for a linear subspace U of a K-vector space $(V, +, m)$, we can restrict $+$ to a map $U \times U \to U$ and m to a map $K \times U \to U$, such that U together with these restrictions becomes itself a vector space over K.

Example 6.7. (1) If K is a field and V is a vector space over K, then $\{0\}$ and V are linear subspaces of V. These are called *the trivial subspaces of V*.

(2) Let K be a field, V be a vector space over K and $v \in V$. Then

$$K \cdot v := \{\lambda v \in K^n \mid \lambda \in K, v \in K^n\}$$

is a linear subspace of V. Property (ii) of a linear subspace is obvious, while for property (i) we observe that $\lambda_1 v + \lambda_2 v = (\lambda_1 + \lambda_2) \cdot v \in K \cdot v$ for all $\lambda_1, \lambda_2 \in K$. For example, for $V = K^n$, V_i is a linear subspace of K^n for each $i \in \{1, 2, \ldots, n\}$, where

$$V_i = \{(0, \ldots, 0, \underbrace{x}_{i\text{-th entry}}, 0, \ldots, 0) \in \mathbb{R}^n \mid x \in \mathbb{R}\} = \mathbb{R} \cdot e_i.$$

Here, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in K^n$ with 1 being in the i-th entry.

(3) Linear subspaces are often defined in terms of linear equations, as we will see in detail below. For example, for $V = \mathbb{R}^2$ as a real vector space

$$U = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 = 0\}$$

is a linear subspace: if $(x_1, x_2), (y_1, y_2) \in U$, then $(x_1 + y_1) + (x_2 + y_2) = 0$, hence $(x_1, x_2) + (y_1, y_2) \in U$. For $\lambda \in \mathbb{R}$, it further holds that $\lambda x_1 + \lambda x_2 = \lambda(x_1 + x_2) = 0$, hence $\lambda(x_1, x_2) \in U$. In contrast to this,

$$U_1 := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = x_1^2\} \quad \text{and} \quad U_2 := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \geq 0\}.$$

are no linear subspaces of $\mathbb{R}^2$, as the reader can easily check.

(4) Consider again the vector space of real sequences $\mathbb{R}^{\mathbb{N}}$. Then it follows from Theorem 2.7 that

$$\{(a_n)_{n\in\mathbb{N}} \mid (a_n)_{n\in\mathbb{N}} \text{ is convergent}\}, \qquad \{(a_n)_{n\in\mathbb{N}} \mid \lim_{n\to\infty} a_n = 0\}$$

are linear subspaces of $\mathbb{R}^{\mathbb{N}}$, while $\{(a_n)_{n\in\mathbb{N}} \mid \lim_{n\to\infty} a_n = 1\}$ is not.

(5) Let I be an interval and let $K = \mathbb{R}$ or $K = \mathbb{C}$. consider the K-vector space K^I. Then:

- $C^0(I, K) := \{f : I \to K \mid f \text{ is continuous}\}$ is a linear subspace of K^I by Theorem 3.5.

- $C^k(I,K) := \{f : I \to K \mid f \text{ is } k \text{ times differentiable and } f^{(k)} \text{ is continuous}\}$ is a linear subspace of K^I for each $k \in \mathbb{N}$. This can be derived from Theorem 4.9.a).
- Given $a, b \in \mathbb{R}$ with $a < b$, $\mathcal{R}([a,b])$ is a linear subspace of $\mathbb{R}^{[a,b]}$ by Theorem 5.11.a) and b).

Theorem 6.8. *Let K be a field and V be a vector space over K. Let I be a set and $(U_i)_{i\in I}$ be a family of linear subspaces of V. Then*

$$U := \bigcap_{i\in I} U_i$$

is a linear subspace of V.

Proof. Since $0 \in U_i$ for all $i \in I$, it holds that $0 \in U$, such that $U \neq \varnothing$. Let $v, w \in U$ and $\lambda \in K$. Then $v, w \in U_i$ for each $i \in I$ and since each U_i is a linear subspace of V it follows that $v + w \in U_i$ and $\lambda v \in U_i$ for each $i \in I$ as well. Hence, $v + w \in U$ and $\lambda v \in U$, such that U is a linear subspace of V. □

Remark 6.9. In contrast to intersections, the *union* of linear subspaces of V is in general not a linear subspace of V. Consider $V = \mathbb{R}^2$ with the two linear subspaces

$$U_1 = \{(x_1, 0) \mid x_1 \in \mathbb{R}\} = \mathbb{R} \cdot (1,0), \qquad U_2 = \{(0, x_2) \mid x_2 \in \mathbb{R}\} = \mathbb{R} \cdot (0,1).$$

Then $(1,0) \in U_1$, $(0,1) \in U_2$, but

$$(1,0) + (0,1) = (1,1) \notin U_1 \cup U_2,$$

such that $U_1 \cup U_2$ is not a linear subspace of $\mathbb{R}^2$.

Definition 6.10. Let K be a field, V be a vector space over K.

a) Let $E \subset V$ be an arbitrary subset. Let $P := \{U \subset V \mid U \text{ is a linear subspace of } V \text{ with } E \subset U\}$. The *span of E* (or *linear hull of E*) is given by

$$\operatorname{span}(E) := \bigcap_{U\in P} U.$$

b) If $U \subset V$ is a linear subspace of V and $S \subset V$ is a subset with $U = \operatorname{span}(S)$, then S is called a *generating set of U*.

Remark 6.11. Let K be a field, V be a vector space over K and $E \subset V$.

(1) By Theorem 6.8, $\operatorname{span}(E)$ is a linear subspace of V.

(2) $\operatorname{span}(E)$ is the *smallest* linear subspace of V which contains E: if U is an arbitrary linear subspace of V with $E \subset U$, then by definition $\operatorname{span}(E) \subset U$.

(3) In the case of $E = \varnothing$, we obtain $\operatorname{span}(\varnothing) = \{0\}$.

(4) Note that by construction, if $E \subset F \subset V$, then $\operatorname{span}(E) \subset \operatorname{span}(F)$.

Theorem 6.12. *Let K be a field, V be a vector space over K and $E \subset V$ with $E \neq \varnothing$. Then*

$$\operatorname{span}(E) = \left\{ \sum_{i=1}^{n} \lambda_i v_i \in V \;\middle|\; n \in \mathbb{N},\ \lambda_1, \ldots, \lambda_n \in K,\ v_1, \ldots, v_n \in E \right\}.$$

Proof. We denote the right-hand side of the equation we want to show by U.
"$\subset$": Let $v = \sum_{i=1}^{n} \lambda_i v_i$, $w = \sum_{j=1}^{m} \mu_j w_j \in U$, where $v_1, \dots, v_n, w_1, \dots, w_m \in E$. Then it immediately follows that

$$v + w = \sum_{i=1}^{n} \lambda_i v_i + \sum_{j=1}^{m} \mu_j w_j \in U.$$

Moreover, for all $\lambda \in K$,

$$\lambda v = \lambda \sum_{i=1}^{n} \lambda_i v_i = \sum_{i=1}^{n} (\lambda \lambda_i) v_i \in U.$$

Hence U is a linear subspace of V. Moreover, $v = 1 \cdot v \in U$ for all $v \in E$, hence $E \subset U$. Thus, by definition of $\operatorname{span}(E)$, we immediately obtain that $\operatorname{span}(E) \subset U$.
"$\supset$": Let $W \subset V$ be an arbitrary linear subspace with $E \subset W$. Then for all $v_1, v_2 \in E$ and $\lambda_1 \in K$ it holds that $v_1 + v_2 \in W$ and $\lambda_1 v_1 \in K$. Iterating these properties, we obtain that $\lambda_1 v_1 + \dots + \lambda_n v_n \in W$ for all $n \in \mathbb{N}$, $v_1, \dots, v_n \in E$ and $\lambda_1, \dots, \lambda_n \in K$, i.e. $U \subset W$. Since $\operatorname{span}(E)$ is a linear subspace of V with $E \subset \operatorname{span}(E)$, it follows that $U \subset \operatorname{span}(E)$. □

Definition 6.13. Let K be a field, V be a vector space over K and let $v_1, \dots, v_n \in V$, where $n \in \mathbb{N}$. Any vector of the form

$$v = \lambda_1 v_1 + \lambda_2 v_2 + \dots + \lambda_n v_n, \qquad \lambda_1, \dots, \lambda_n \in K,$$

it called a *linear combination of* $v_1, \dots, v_n$.

In this terminology Theorem 6.12 says that $\operatorname{span}(E)$ is the set of all linear combinations of vectors in E.

Definition 6.14. Let K be a field, V be a vector space over K and $(U_i)_{i \in I}$ be a family of linear subspaces of V. The *sum of* $(U_i)_{i \in I}$ is given by

$$\sum_{i \in I} U_i := \operatorname{span}\left(\bigcup_{i \in I} U_i \right).$$

If $I = \{1, 2, \dots, n\}$ is finite, then we also write $U_1 + U_2 + \dots + U_n := \sum_{i \in \{1,2,\dots,n\}} U_i$.

Theorem 6.15. *Let K be a field and V be a vector space over K. Let $n \in \mathbb{N}$ and let $U_1, \dots, U_n$ be linear subspaces of V. Then*

$$U_1 + U_2 + \dots + U_n = \{ v_1 + v_2 + \dots + v_n \in V \mid v_i \in U_i \; \forall i \in \{1, 2, \dots, n\} \}.$$

Proof. Let U denote the set on the right-hand side of the equation. Let $v := v_1 + \dots + v_n, w := w_1 + \dots + w_n \in U$, where $v_i, w_i \in U_i$ for all $i \in \{1, 2, \dots, n\}$. Since the U_i are linear subspaces, $v_i + w_i \in U_i$ for each i, hence

$$v + w = (v_1 + w_1) + (v_2 + w_2) + \dots (v_n + w_n) \in U.$$

Let $\lambda \in K$. Since the U_i are linear subspaces, $\lambda v_i \in U_i$ for each i and thus

$$\lambda v = \lambda(v_1 + \dots + v_n) = \lambda v_1 + \dots + \lambda v_n \in U.$$

Thus, U is a linear subspace of V. With this observation, we prove both inclusions.

"$\subset$": Since obviously $U_i \subset U$ for all $i \in \{1, 2, \ldots, n\}$ it follows that

$$U_1 + \cdots + U_n = \operatorname{span}(U_1 \cup \cdots \cup U_n) \subset U.$$

"$\supset$": If $W \subset V$ is an arbitrary linear subspace of V with $U_i \subset V$ for all $i \in \{1, 2, \ldots, n\}$, then it particularly holds for all $v, w \in U_1 \cup \cdots \cup U_n$ that $v + w \in W$. Iterating this observation shows that $v_1 + \cdots + v_n \in W$ for all $v_i \in U_i$, $i \in \{1, 2, \ldots, n\}$ and thus $U \subset W$. In particular, this implies that $U \subset \operatorname{span}(E)$. □

6.2 Bases and dimensions of vector spaces

In physics one often considers configurations of mechanical systems as points in $\mathbb{R}^n$ for some suitable $n \in \mathbb{N}$, e.g. the position and momentum of an object as a point in the *phase space* of a mechanical system in Hamiltonian mechanics. Here, we use the notion of "n-dimensional space" as intuitively clear, the dimension is roughly seen as the number of coordinate axes. In order to obtain a formal definition of dimension, we need to study bases of vector spaces. Intuitively, a basis of a vector space V is a subset of V that contains precisely the "coordinate directions" of V in the sense that all elements of V can be expressed in terms of these coordinate directions as we shall formalize below.

Throughout this section, let K be a field and let V be a vector space over K.

Definition 6.16. A subset $E \subset V$ is *linearly independent* if the following holds: if $v_1, \ldots, v_n \in E$ and $\lambda_1, \ldots, \lambda_n \in K$, where $n \in \mathbb{N}$, are such that

$$\sum_{i=1}^{n} \lambda_i v_i = 0,$$

then $\lambda_1 = 0$, $\lambda_2 = 0$, …, $\lambda_n = 0$. A subset of V which is not linearly independent is called *linearly dependent*.

Example 6.17. (1) Let $n \in \mathbb{N}$ and let again $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in K^n$, where 1 is in the k-th entry of e_i, for all $i \in \{1, 2, \ldots, n\}$. Then $\{e_1, \ldots, e_n\} \subset K^n$ is linearly independent: given $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$, it holds that

$$\lambda_1 e_1 + \ldots + \lambda_n e_n = 0 \quad \Leftrightarrow \quad (\lambda_1, \lambda_2, \ldots, \lambda_n) = 0 \quad \Leftrightarrow \quad \lambda_1 = \lambda_2 = \ldots = \lambda_n = 0.$$

(2) $\{(1, 0, 0), (2, 0, 0)\}$ is linearly dependent in $\mathbb{R}^3$, since

$$2 \cdot (1, 0, 0) + (-1) \cdot (2, 0, 0) = 0.$$

(3) Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$, $f_1(x) = x$, $f_2(x) = x^2$. Then $\{f_1, f_2\}$ is linearly independent in $C^0(\mathbb{R}, \mathbb{R})$: if $\lambda_1, \lambda_2 \in \mathbb{R}$ satisfy $\lambda_1 f_1 + \lambda_2 f_2$, then

$$0 = \lambda_1 f_(x) + \lambda_2 f_2(x) = \lambda_1 x + \lambda_2 x^2 = x(\lambda_1 + \lambda_2 x) \qquad \forall x \in \mathbb{R},$$

whose only possibility is $\lambda_1 = \lambda_2 = 0$.

Remark 6.18. We summarize some general observations on the linear (in)dependence of subsets of V.

(1) Let $v \in V$. Then $\{v\}$ is linearly independent if and only if $v \neq 0$. Since $\lambda \cdot 0 = 0$ for each $\lambda \in K$, $\{0\}$ is linearly dependent, while the other direction is obvious.

(2) $E = \varnothing$ is linearly independent, since there is no linear combination of elements of E which equals 0.

(3) If $E \subset V$ is linearly dependent, then any $F \subset V$ with $E \subset F$ is linearly dependent as well. With (1) this shows that if $E \subset V$ is linearly independent, then $0 \notin E$.

(4) If $E \subset V$ is linearly independent, then any $F \subset V$ with $F \subset E$ is linearly independent as well. This is obvious.

Theorem 6.19. *Let $E \subset V$ be a subset. The E is linearly independent if and only if the following holds: For any $v \in \operatorname{span}(E)$ there are* **unique** *$n \in \mathbb{N}$, $v_1, \ldots, v_n \in E$ and $\lambda_1, \ldots, \lambda_n \in K$, such that*

$$v = \sum_{i=1}^{n} \lambda_i v_i.$$

Proof. "⇒": Let $v \in \operatorname{span}(E)$. Then we can write $v = \sum_{e \in E} \lambda_e e$, where $\lambda_e = 0$ for all but finitely many $e \in E$. Assume there are numbers $\mu_e \in K$, $e \in E$ with $\mu_e = 0$ for all but finitely many $e \in E$, such that $v = \sum_{e \in E} \mu_e e$ as well. Then

$$0 = v - v = \sum_{e \in E} \lambda_e e - \sum_{e \in E} \mu_e e = \sum_{e \in E} (\lambda_e - \mu_e) e.$$

Since E is linearly independent, this is equivalent to $\mu_e = \lambda_e$ for all $e \in E$. Hence, the representation of v as a linear combination of vector in E is unique.
"⇐": Let $n \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_n \in K$, $v_1, \ldots, v_n \in E$, such that $\sum_{i=1}^{n} \lambda_i v_i = 0$. Since $0 = \sum_{i=1}^{n} 0 \cdot v_i$, it follows by assumption that $\lambda_i = 0$ for all $i \in \{1, 2, \ldots, n\}$. Hence, E is linearly independent. □

Another more concise, but more abstract criterion for linear independence is the following.

Theorem 6.20. *Let $E \subset V$. The following statements are equivalent:*

(i) E is linearly independent.

(ii) For all $F \subset E$ with $F \neq E$ it holds that $\operatorname{span}(F) \neq \operatorname{span}(E)$.

Proof. (i) ⇒ (ii): Let $F \subset E$ with $F \neq E$. Assume there exists $v \in E \setminus F$, but $v \in \operatorname{span}(F)$. Then by Theorem 6.12 there are $v_1, \ldots, v_n \in F$, $\lambda_1, \ldots, \lambda_n \in K$, such that

$$v = \sum_{i=1}^{n} v_i \quad \Leftrightarrow \quad \sum_{i=1}^{n} \lambda_i v_i - 1 \cdot v = 0.$$

But $v \neq 0$ since $v \notin F$. Since $v_1, \ldots, v_n, v \in E$, this is a contradiction to the linear independence of E. Thus, $v \notin \operatorname{span}(F)$ for all $v \in E \setminus F$, which shows that $\operatorname{span}(F) \neq \operatorname{span}(E)$.
(ii) ⇒ (i): Suppose E was linearly dependent. Then there exist $\lambda_e \in K$, $e \in E$, with $\sum_{e \in E} \lambda_e e = 0$ and with $\lambda_{e_0} \neq 0$ for some $e_0 \in E$. But then:

$$\sum_{e \in E} \lambda_e e = 0 \quad \Leftrightarrow \quad \lambda_{e_0} e_0 = - \sum_{e \in E \setminus \{e_0\}} \lambda_e e \quad \Leftrightarrow \quad e_0 = -\frac{1}{\lambda_{e_0}} \sum_{e \in E \setminus \{e_0\}} \lambda_e e \in \operatorname{span}(E \setminus \{e_0\}).$$

Consequently, $\operatorname{span}(E) = \operatorname{span}(E \setminus \{e_0\})$, which contradicts the assumption. Hence, E is linearly independent. □

Definition 6.21. A subset $B \subset V$ is called *a basis of* V if $\operatorname{span}(B) = V$ and if B is linearly independent.

Example 6.22. (1) $\varnothing$ is a basis of $\{0\}$.

(2) In the notation of Example 6.17.(1), $\{e_1, \dots, e_n\}$ is a basis of K^n. This basis is called *the canonical basis* or *standard basis of* K^n.

(3) $\{(1,2),(0,1)\}$ is a basis of $\mathbb{R}^2$: We can write any $(x_1, x_2) \in \mathbb{R}^2$ as

$$(x_1, x_2) = x_1 \cdot (1,2) + (x_2 - 2x_1) \cdot (0,1),$$

hence $\operatorname{span}(\{(1,2),(0,1)\}) = \mathbb{R}^2$. Moreover, $\{(1,2),(0,1)\}$ is linearly independent, since

$$\lambda_1 \cdot (1,2) + \lambda_2 \cdot (0,1) = 0 \quad \Leftrightarrow \quad (\lambda_1, 2\lambda_1 + \lambda_2) = (0,0) \quad \Leftrightarrow \quad \lambda_1 = \lambda_2 = 0.$$

(4) Let $K[x] := \{f : K \to K \mid f \text{ is a real polynomial in the variable } x\}$. Then K is easily seen to be a linear subspace of K^K, hence a real vector space. Then

$$B = \{1, x, x^2, x^3, \dots, x^n, \dots\}$$

is a basis of $K[x]$. It is obvious that $\operatorname{span}(B) = K[x]$, while the linear independence of B follows from some elementary results about polynomials that we shall omit. Elements of B are also called *monomials*.

(5) Let X be an infinite set and consider the real vector space $\mathbb{R}^X$. For each $x \in X$ define

$$\delta_x : X \to \mathbb{R}, \qquad \delta_x(y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

Then $\{\delta_x \in \mathbb{R}^X \mid x \in X\}$ is *not* a basis of $\mathbb{R}^X$. For example, the function $f : X \to \mathbb{R}$, $f(x) = 1$ for all $x \in X$ can *not* be written as a linear combination of the δ_x, since linear combinations are by definition *finite* sums. However, since $\{\delta_x \in \mathbb{R}^X \mid x \in X\}$ is easily seen to be linearly independent, it is a basis of the linear subspace

$$U := \operatorname{span}\Big(\{\delta_x \in \mathbb{R}^X \mid x \in X\}\Big) = \{f : X \to \mathbb{R} \mid f(x) = 0 \text{ for all but finitely many } x \in X\}.$$

Lemma 6.23. *Let $E \subset V$ be linearly independent with $\operatorname{span}(E) \neq V$. Then for all $v \in V \setminus \operatorname{span}(E)$, the set $E \cup \{v\}$ is linearly independent.*

Proof. Let $\lambda_e \in K$ for each $e \in E$, such that $\lambda_e = 0$ for all but finitely many e and let $\mu \in K$ with

$$\sum_{e \in E} \lambda_e e + \mu v = 0.$$

If $\mu \neq 0$, then $v = -\sum_{e \in E} \frac{\lambda_e}{\mu} e \in \operatorname{span}(E)$, which contradicts the assumption, so $\mu = 0$, which implies $\sum_{e \in E} \lambda_e e = 0$. Since E is linearly independent, this yields $\lambda_e = 0$ for each $e \in E$. Thus, $E \cup \{v\}$ is linearly independent. □

Theorem 6.24. *Let $B \subset V$. The following statements are equivalent:*

(i) B is a basis of V.

(ii) B *is a* minimal generating set *of* V*, i.e.* $\operatorname{span}(B) = V$*, but* $\operatorname{span}(B') \neq V$ *for all* $B' \subset B$ *with* $B' \neq B$*.*

(iii) B *is a* maximal linearly independent subset *of* V*, i.e.* B *is linearly independent and every* $E \subset V$ *with* $B \subset E$ *and* $B \neq E$ *is linearly dependent.*

Proof. The equivalence of (i) and (ii) is a special case of Theorem 6.20, since $V = \operatorname{span}(B)$, so it suffices to show that (i) and (iii) are equivalent.

(i) $\Rightarrow$ (iii): Let $E \subset V$ with $B \subset E$, but $B \neq E$ and let $v \in E \setminus B$. Since B is a basis of V, there are $v_1, \dots, v_n \in B$, $\lambda_1, \dots, \lambda_n \in K$, such that

$$v = \sum_{i=1}^{n} \lambda_i v_i \quad \Leftrightarrow \quad \sum_{i=1}^{n} \lambda_i v_i - 1 \cdot v = 0.$$

Since $v_1, \dots, v_n, v \in E$, this shows that E is linearly dependent, which shows the claim.
(iii) $\Rightarrow$ (i): Let B be a maximal linearly independent subset of V and suppose B was not a basis of V, i.e. that $\operatorname{span}(B) \neq V$. Then for $v \in V \setminus \operatorname{span}(B)$ it follows from Lemma 6.23 that $B \cup \{v\}$ is linearly independent. This contradicts the maximality property of B, hence B is a basis of V. □

Theorem 6.25. *Assume that* V *has a finite generating set* E *and let* $A \subset E$ *be linearly independent. Then there exists a basis* B *of* V *with* $A \subset B \subset E$*.*

Proof. Let

$$\mathcal{F}(A, E) := \{F \subset E \mid A \subset F \text{ and } F \text{ is linearly independent}\}.$$

First of all, $\mathcal{F}(A, E) \neq \varnothing$ since $A \in \mathcal{F}(A, E)$.
Let $F_1 \in \mathcal{F}(A, E)$. be given. If F_1 is not a basis of V, i.e. if $\operatorname{span}(F_1) \neq V = \operatorname{span}(E)$, then by Lemma 6.23 there exists $F_2 \in F(A, E)$ with $F_1 \subset F_2$ and $F_1 \neq F_2$. If F_2 is not a basis of V, then analogously there exists $F_3 \in F(A, E)$ with $F_2 \subset F_3$ and $F_2 \neq F_3$. Defining $F_3, F_4, \dots \in \mathcal{F}(A, E)$ by proceeding iteratively, since E is finite, there must be a $k \in \{1, 2, \dots, n\}$ with $\operatorname{span}(F_k) = \operatorname{span}(E) = V$. Then F_k is a basis of V. □

Remark 6.26. In fact, Theorem 6.25 holds true for *every* vector space V and the assumption of a finite generating set can be dropped entirely. However, a proof of the general statement (which we will not need in this course) requires the use of the famous and infamous Zorn's lemma about partially ordered sets from set theory.[2]

We next summarize several important consequences of Theorem 6.25. Since they follow straightaway from that theorem, we omit their proofs.

Corollary 6.27. *Assume that* V *has a finite generating set. Then:*

a) V *has a basis.*

b) If $F \subset V$ *is linearly independent, then there exists a basis* B *of* V *with* $F \subset B$*.*

c) Every finite generating set of V *contains a basis.*

Another important consequence is discussed separately in the following theorem.

[2] https://en.wikipedia.org/wiki/Zorn%27s_lemma

Theorem 6.28 (Basis extension theorem)**.** *Assume that V has a finite generating set E and let $A \subset E$ be finite and linearly independent. Then there exists $B' \subset E$, such that $A \cup B'$ is a basis of V.*

Proof. Since A is finite, $A \cup E$ is another finite generating set of V. Thus, by Theorem 6.25, there exists a basis B of V with $A \subset B \subset A \cup E$. The claim thus follows with $B' := B \setminus A$. □

In the remainder of this section we want to take a closer look on bases of V in the case that V has a finite generating set. We shall see that this condition on V is equivalent to an existence condition for bases.

Lemma 6.29 (Exchange lemma)**.** *Let B be a basis of V and E be a generating set of V. Then for each $b \in B$ there exists $e \in E$, such that*

$$(B \setminus \{b\}) \cup \{e\}$$

is a basis of V.

Proof. Let $b \in B$. By Theorem 6.20, $\operatorname{span}(B \setminus \{b\}) \neq \operatorname{span}(B) = V$, hence $B \setminus \{b\}$ is not a generating set of V. Thus, there exists $e \in E$ with $e \notin \operatorname{span}(B \setminus \{b\})$ and by Lemma 6.23, the set $B' := (B \setminus \{b\}) \cup \{e\}$ is linearly independent in V.
To show that B' is a generating set of V, it suffices to show that $b \in \operatorname{span}(B')$, since then $\operatorname{span}(B') = \operatorname{span}(B) = V$. Since $e \notin \operatorname{span}(B \setminus \{b\})$, but B is a basis of V, there are $v_1, \dots, v_n \in B \setminus \{b\}$ and $\lambda_1, \dots, \lambda_n, \mu \in K$ with $\mu \neq 0$, such that

$$e = \sum_{i=1}^{n} \lambda_i v_i + \mu b \quad \Leftrightarrow \quad b = \frac{1}{\mu} e - \sum_{i=1}^{n} \frac{\lambda_i}{\mu} v_i \in \operatorname{span}((B \setminus \{b\}) \cup \{e\}) = \operatorname{span}(B'),$$

which implies $V = \operatorname{span}(B')$, such that B' is a basis of V. □

Given a finite set M we let $|M|$ denote the number of elements of M.

Theorem 6.30 (Steinitz's exchange theorem)**.** *Assume that V has a finite generating set E and let B be a basis of V. Then there exists a subset $B' \subset E$, such that B' is a basis of V and $|B'| = |B|$. In particular, $|B| \leq |E|$.*

Proof. Let $b_1 \in B$. By Lemma 6.29 there exists $e_1 \in E$, such that

$$B_1 := (B \setminus \{b_1\}) \cup \{e_1\}$$

is a basis of E. Let $b_2 \in B \setminus \{b_1\}$. Then again by Lemma 6.29 there exists $e_2 \in E$, such that

$$B_2 := (B_1 \setminus \{b_2\}) \cup \{e_2\} = (B \setminus \{b_1, b_2\}) \cup \{e_1, e_2\}.$$

is a basis of V. Further iterating this procedure shows that for any $\{b_1, \dots, b_k\} \subset B$, $k \in \mathbb{N}$, there exists $\{e_1, \dots, e_k\} \subset E$, such that

$$B_k := (B \setminus \{b_1, \dots, b_k\}) \cup \{e_1, \dots, e_k\}$$

is a basis of V. Assume that $|B| > |E| =: n$. Then there is $\{b_1, \dots, b_n\} \subset B$, such that

$$B_n = (\underbrace{B \setminus \{b_1, \dots, b_n\}}_{\neq \varnothing}) \cup E$$

is a basis of V. But then, by Theorem 6.24, $\operatorname{span}(E) \neq \operatorname{span}(B_n) = V$, which is a contradiction. Thus, $r := |B| \leq |E|$ and so, after r iterations, we obtain $B' = \{e_1, \dots, e_r\} \subset E$, such that B_r is a basis of V with $|B'| = r$. □

The most important consequence of Theorem 6.30 is the following corollary.

Corollary 6.31. *Assume that V has a finite basis B. Then every basis of V is finite and it holds that $|B'| = |B|$ for all bases B' of V.*

Proof. Let B' be another basis of V. Since B' is linearly independent, it holds by Theorem 6.30 that $|B'| \leq |B|$, in particular that B' is finite. Another application of Theorem 6.30 with the roles of B and B' reversed shows that $|B| \leq |B'|$. □

The previous corollary motivates the following condition.

Definition 6.32. We call V *finite-dimensional* if it has a finite basis. The number of elements of any (hence every) basis of V is called *the dimension of V* and denoted by $\dim V$.
If V does not have a finite basis, then we call it *infinite-dimensional*.

If V is finite-dimensional, then Theorem 6.30 provides us with a simple criterion for a linearly independent subset of V to be a basis.

Corollary 6.33. *Let V be finite-dimensional, and let $F \subset V$ be linearly independent. Then $|F| \leq \dim V$ and F is a basis of V if and only if $|F| = \dim V$.*

Proof. By the basis extension theorem, there exists $A \subset V$, such that $F \cup A$ is a basis of V, hence $|F| \leq |F \cup A| = \dim V$ by Corollary 6.31. If F is a basis of V, then the same corollary yields $|F| = \dim V$. Conversely, if $|F| = \dim V$, then it follows that $|F| = |F \cup A|$, i.e. $A \subset F$, which implies that F is a basis of V. □

Remark 6.34. Note that V is finite-dimensional if and only if it has a finite generating set. While "⇒" is trivial, the other implication follows from Corollary 6.27.c).

Example 6.35. (1) $\dim K^n = n$ for each $n \in \mathbb{N}$, since the canonical basis $\{e_1, \ldots, e_n\}$ of K^n has precisely n elements.

(2) $\dim\{0\} = 0$, since the empty set is a basis of $\{0\}$.

(3) $\dim C^k(I, \mathbb{R}) = +\infty$ for each interval I and each $k \in \mathbb{N}_0$. Since $\mathbb{R}[x] \subset C^k(I, \mathbb{R})$ is a linear subspace which contains the infinite linearly independent set $\{x^n \mid n \in \mathbb{N}_0\}$, this follows from Corollary 6.33.

END OF MATHEMATICS FOR PHYSICISTS 1

6.3 Complements and direct sums of linear subspaces

In this section we want to formalize the notion that a vector space can be "decomposed" into suitable linear subspaces. We shall see applications of such decompositions later in this course.

Throughout this section let K be a field and let V be a vector space over K.

Definition 6.36. Let $U \subset V$ be a linear subspace. A linear subspace $W \subset V$ is called a *complement of U in V* if

$$V = U + W \qquad \text{and} \qquad U \cap W = \{0\}.$$

If W is a complement of U, then we also write

$$V = U \oplus W$$

and call V *the direct sum of U and V*.

With regards to the introduction of this section, a direct sum $V = U \oplus W$ is interpreted as a "decomposition" of V into the linear subspaces U and W.

Example 6.37. (1) V is a complement of $\{0\}$ in V and V is a complement of $\{0\}$ in V.

(2) Let $U_1 = \{(x,0) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ and $U_2 = \{(0,y) \in \mathbb{R}^2 \mid y \in \mathbb{R}\}$. Then $\mathbb{R}^2 = U_1 \oplus U_2$:
Obviously, $U_1 \cap U_2 = \{0\}$ and $(x,y) = (x,0) + (0,y) \in U_1 + U_2$ for all $(x,y) \in \mathbb{R}^2$, hence $U_1 + U_2 = \mathbb{R}^2$.

(3) Let $U_3 = \{(y,y) \in \mathbb{R}^2 \mid y \in \mathbb{R}\}$. Then $\mathbb{R}^2 = U_1 \oplus U_3$:
$U_1 \cap U_3 = \{0\}$ is obvious and $(x,y) = (x-y,0) + (y,y) \in U_1 + U_3$ for all $(x,y) \in \mathbb{R}^2$.
This example shows that complements of linear subspaces are in general not unique.

(4) Let $W_1 := \{(x,y,0) \in \mathbb{R}^3 \mid (x,y) \in \mathbb{R}^2\}$, $W_2 := \{(0,y,z) \in \mathbb{R}^3 \mid (y,z) \in \mathbb{R}^2\}$. Then $\mathbb{R}^3 = W_1 + W_2$, but $\mathbb{R}^3 \neq W_1 \oplus W_2$, since $W_1 \cap W_2 = \{(0,y,0) \mid y \in \mathbb{R}\} \supsetneq \{0\}$.

(5) Let $A_1 = \{f \in C^0(\mathbb{R},\mathbb{R}) \mid f(0) = 0\}$ and let $A_2 := \{c_a \in C^0(\mathbb{R},\mathbb{R}) \mid c_a(x) = a \ \forall x \in \mathbb{R}, a \in \mathbb{R}\}$ be the set of constant functions.. One checks without difficulties that both A_1 and A_2 are linear subspaces of $C^0(\mathbb{R},\mathbb{R})$ and obviously, $A_1 \cap A_2 = \{c_0\}$, which is the zero vector in $C^0(\mathbb{R},\mathbb{R})$. Moreover, for each $f \in C^0(\mathbb{R},\mathbb{R})$ it holds that

$$f = \underbrace{(f - c_{f(0)})}_{\in A_1} + c_{f(0)} \in A_1 + A_2,$$

hence $C^0(\mathbb{R},\mathbb{R}) = A_1 \oplus A_2$.

If a vector space is given as the direct sum of linear subspaces, this has an interesting consequence for the individual vectors in this vector space.

Theorem 6.38. *Let $U_1, U_2 \subset V$ be linear subspaces with $V = U_1 \oplus U_2$. Then for all $v \in V$ there are* unique *$u_1 \in U_1$ and $u_2 \in U_2$ with*

$$v = u_1 + u_2.$$

Proof. Let $v \in V$. Since $V = U_1 + U_2$, the existence of such u_1 and u_2 follows from Theorem 6.15. Let $u_1, v_1 \in U_1$ and $u_2, v_2 \in U_2$ be such that $v = u_1 + u_2 = v_1 + v_2$. Then $u_1 - v_1 = u_2 - v_2$. Since $u_1 - v_1 \in U_1$, $u_2 - v_2 \in U_2$, this shows that $u_1 - v_1 \in U_1 \cap U_2 = \{0\}$, hence $u_1 - v_1 = u_2 - v_2 = 0$, which yields $u_1 = v_1$ and $u_2 = v_2$. So u_1 and u_2 are indeed unique. □

So far we have seen examples of complements of linear subspaces, but we have not discussed the existence of complements. For finite-dimensional vector spaces this question is answered by the following result.

Theorem 6.39. *Let V be finite-dimensional. Then every linear subspace of V has a complement in V.*

Proof. Let $U \subset V$ be a linear subspace. One derives from Corollary 6.33 that U is finite-dimensional, thus has a basis $B_1 \subset U$ by Corollary 6.27.a). By the basis extension theorem, there exists $B_2 \subset V$ with $B_1 \cap B_2 = \varnothing$, such that $B_1 \cup B_2$ is a basis of V. Put $W := \operatorname{span}(B_2)$. We want to show that $V \overset{!}{=} U \oplus W$. Let $v \in U \cap W$. Then there are $\lambda_b, \mu_c \in K$ for each $b \in B_1$ and $c \in B_1$, such that

$$v = \sum_{b \in B_1} \lambda_b \cdot b = \sum_{c \in B_2} \mu_c \cdot c \quad \Rightarrow \quad \sum_{b \in B_1} \lambda_b \cdot b + \sum_{c \in B_2} (-\mu_c) \cdot c = 0.$$

But since $B_1 \cup B_2$ is linearly independent, this yields $\lambda_b = 0$ for all $b \in B_1$ and thus $v = 0$. Hence, $U \cap W = \{0\}$. Since $B_1 \cup B_2$ is a basis of V, we further obtain that

$$V \supset U + W = \operatorname{span}(U \cup W) \overset{\text{Remark 6.11.(4)}}{\supset} \operatorname{span}(B_1 \cup B_2) = V.$$

This shows that $V = U + W$, so we have shown that W is indeed a complement of U. □

To conclude this chapter, we will show a useful formula for the dimension of a (not necessarily direct) sum of linear subspaces in finite-dimensional vector spaces.

Theorem 6.40 (dimension formula[3])**.** *Let V be finite-dimensional and let $U_1, U_2 \subset V$ be linear subspaces. Then*

$$\dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).$$

Proof. By Theorem 6.8, $U_1 \cap U_2$ is a linear subspace of V, it has a (finite) basis $B_0 = \{b_1, \dots, b_r\}$, where $r = \dim(U_1 \cap U_2)$. Since $U_1 \cap U_2$ is a linear subspace both of U_1 and U_2, we can apply the basis extension theorem and extend B_0 to

$$\begin{aligned} \text{a basis } B_1 &= \{b_1, \dots, b_r, v_1, \dots, v_n\} \text{ of } U_1 \\ \text{and a basis } B_2 &= \{b_1, \dots, b_r, w_1, \dots, w_m\} \text{ of } U_2, \end{aligned}$$

for suitable $m, n \in \mathbb{N}$. Then the set $B_1 \cup B_2 = \{b_1, \dots, b_r, v_1, \dots, v_n, w_1, \dots, w_m\}$ has two important properties:

(1) *$B_1 \cup B_2$ is a generating set of $U_1 + U_2$.*

If $v \in U_1 + U_2$, then there are $u_1 \in U_1$ and $u_2 \in U_2$ with $v = u_1 + u_2$. Since B_i is a basis of U_i for $i \in \{1, 2\}$, there are $\lambda_1, \dots, \lambda_{r+n}, \mu_1, \dots, \mu_{r+m} \in K$ with

$$\begin{aligned} u_1 = \sum_{i=1}^{r} \lambda_i b_i + \sum_{j=1}^{n} \lambda_{r+j} v_j, \quad u_2 = \sum_{i=1}^{r} \mu_i b_i + \sum_{j=1}^{m} \mu_{r+j} w_j \\ \Rightarrow \; v = \sum_{i=1}^{r} (\lambda_i + \mu_i) b_i + \sum_{j=1}^{n} \lambda_{r+j} v_j + \sum_{j=1}^{m} \mu_{r+j} w_j \in \operatorname{span}(B_1 \cup B_2). \end{aligned}$$

[3]Unfortunately, there are several results in linear algebra that go by the name "dimension formula". We will already see another dimension formula in the next chapter, but it should always be clear which one we are referring to.

(2) *$B_1 \cup B_2$ is linearly independent.*

Assume that $\lambda_1, \dots, \lambda_r, \mu_1, \dots, \mu_n, \nu_1, \dots, \nu_m \in K$ are given such that

$$\sum_{i=1}^{r} \lambda_i b_i + \sum_{j=1}^{n} \mu_j v_j + \sum_{k=1}^{m} \nu_k w_k = 0. \tag{6.1}$$

Then

$$\sum_{k=1}^{m} \nu_k w_k = -\sum_{i=1}^{r} \lambda_i b_i - \sum_{j=1}^{n} \mu_j v_j \in \operatorname{span}(B_1) = U_1,$$

while $\sum_{k=1}^{m} \nu_k w_k \in U_2$ by definition. Hence, $\sum_{k=1}^{m} \nu_k w_k \in U_1 \cap U_2 = \operatorname{span}(B_0)$, so there are $\xi_1, \dots, \xi_r \in K$, such that $\sum_{k=1}^{m} \nu_k w_k = \sum_{i=1}^{r} \xi_i b_i$. But since B_2 is linearly independent, it follows that $\xi_1 = \cdots = \xi_r = \nu_1 = \cdots = \nu_k = 0$. Inserting this into (6.1) yields

$$\sum_{i=1}^{r} \lambda_i b_i + \sum_{j=1}^{n} \mu_j v_j = 0.$$

Since B_1 is linearly independent, it follows that $\lambda_1 = \cdots = \lambda_r = \mu_1 = \cdots = \mu_n = 0$. Thus, (6.1) has only the trivial solution, such that $B_1 \cup B_2$ is linearly independent.

We have shown that $B_1 \cup B_2$ is a basis of $U_1 + U_2$ and derive:

$$\dim(U_1 + U_2) = r + n + m = (r + n) + (r + m) - r = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).$$

□

Corollary 6.41. *Let V be finite-dimensional and let $U_1, U_2 \subset V$ be linear subspaces of V with $V = U_1 + U_2$. Then $V = U_1 \oplus U_2$ if and only if*

$$\dim V = \dim U_1 + \dim U_2.$$

Proof. "⇒": This is the special case of $\dim(U \cap V) = 0$ of Theorem 6.40.
"⇐": If $\dim V = \dim U_1 + \dim U_2$, then by the dimension formula

$$\dim(U_1 \cap U_2) = \dim U_1 + \dim U_2 - \dim V = 0 \quad \Rightarrow \quad U_1 \cap U_2 = \{0\},$$

which we needed to show. □

Remark 6.42. If V is finite-dimensional and $U \subset V$ is a linear subspace, then Corollary 6.41 particularly implies that all complements of U have the same dimension $\dim V - \dim U$.

Chapter 7

Linear maps

7.1 Definition and first properties of linear maps

In this section we want to introduce the notion of a *linear map* between vector spaces. Roughly speaking, a linear map is a map that is "compatible" with the vector space structures of its domain and its codomain. With respect to additions and multiplications by scalars, a linear map maps sums to sums and scalar multiples to scalar multiples as we shall formally define in the next definition.

Throughout this section, let K be a field.

Definition 7.1. Let V and W be vector spaces over K. A map $f : V \to W$ is called a *linear map* or a *homomorphism* if

(i) $f(v_1 + v_2) = f(v_1) + f(v_2)$ for all $v_1, v_2 \in V$,

(ii) $f(\lambda \cdot v) = \lambda \cdot f(v)$ for all $\lambda \in K$ and $v \in V$.

The set of all linear maps from V to W is denoted by $L(V, W)$ or $\mathrm{Hom}(V, W)$.

Remark 7.2. Let V and W be vector spaces over K and let $f : V \to W$ be a linear map.

(1) By iterating properties (i) and (ii) of Definition 7.1, one shows for all $n \in \mathbb{N}$ that

$$f\Big(\sum_{i=1}^{n} \lambda_i v_i\Big) = \sum_{i=1}^{n} \lambda_i f(v_i) \qquad \forall \lambda_1, \ldots, \lambda_n \in K,\ v_1, \ldots, v_n \in V.$$

(2) Every linear map satisfies $f(0) = 0$: this follows from property (i), since

$$f(0) = f(0+0) \overset{\text{(i)}}{=} f(0) + f(0) \quad \Rightarrow \quad f(0) = 0.$$

(3) In the case that V and W are infinite-dimensional, f is sometimes also called an *operator*.

Example 7.3. (1) Let V and W be vector spaces over K. The zero map $z : V \to W$, $z(v) = 0$, and the *identity map* $\mathrm{id}_V : V \to V$, $\mathrm{id}_V(v) = v$, are linear maps.

(2) Let V be a vector space over K and $a \in K$. Then $f : V \to V$, $f(v) = av$, is linear:

$$\begin{aligned} f(v+w) &= a(v+w) = av + aw = f(v) + f(w) \qquad \forall v, w \in V, \\ f(\lambda v) &= a(\lambda v) = (a\lambda)v = (\lambda a)v = \lambda(av) = \lambda f(v) \qquad \forall \lambda \in K,\ v \in V. \end{aligned}$$

is linear, where we used the field properties of K and the vector space properties of V.

(3) Consider $\mathbb{R}^2$ as a vector space over $\mathbb{R}$ and let

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \qquad f(x, y) = (2x + y, x - 5y).$$

Then f is linear: for each $\lambda \in \mathbb{R}$ we compute that

$$\begin{aligned} f(\lambda(x, y)) = f(\lambda x, \lambda y) &= (2\lambda x + \lambda y, \lambda x - 5\lambda y) \\ &= (\lambda(2x + y), \lambda(x - 5y)) \\ &= \lambda(2x + y, x - 5y) = \lambda f(x, y), \end{aligned}$$

while for all $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$ we compute that

$$\begin{aligned} f((x_1, y_1) + (x_2, y_2)) &= f(x_1 + x_2, y_1 + y_2) \\ &= (2(x_1 + x_2) + y_1 + y_2, x_1 + x_2 - 5(y_1 + y_2)) \\ &= (2x_1 + y_1 + 2x_2 + y_2, x_1 - 5y_1 + x_2 - 5y_2) \\ &= (2x_1 + y_1, x_1 - 5y_1) + (2x_2 + y_2, x_2 - 5y_2) = f(x_1, y_1) + f(x_2, y_2). \end{aligned}$$

(4) Let $I \subset \mathbb{R}$ be an open interval, let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $C^1(I, K)$ and $C^0(I, K)$ be given as in Example 6.7.(5) as vector spaces over K. The map $D : C^1(I, K) \to C^0(I, K)$, $D(f) = f'$, is linear. This follows directly from Theorem 4.9.a).

(5) Let $a, b \in \mathbb{R}$ with $a < b$. The map $I : \mathcal{R}([a, b]) \to \mathbb{R}$, $I(f) = \int_a^b f(x)\ dx$, is linear. This follows directly from Theorem 5.11.a) and b).

(6) A counterexample: Given $v_0 \in V$ with $v_0 \neq 0$, the map $f : V \to W$, $f(v) = v + v_0$ is *not* linear: for all $v, w \in V$ we compute

$$f(v) + f(w) = v + w + 2v_0 \neq v + w + v_0 = f(v + w).$$

In the following theorems, we explore some elementary properties of linear maps.

Theorem 7.4. *Let V and W be linear subspaces over K.*

a) If $f, g : V \to W$ are linear, then $f + g : V \to W$ is linear.

b) If $f : V \to W$ is linear and $\lambda \in K$, then $\lambda f : V \to W$ is linear.

Moreover, together with addition and scalar multiplication of maps, $L(V, W)$ is a vector space over K

Proof. a) Let $v, w \in V$ and $\lambda \in K$. Then, since f and g are linear,

$$\begin{aligned} (f+g)(v+w) &= f(v+w) + g(v+w) \\ &= f(v) + f(w) + g(v) + g(w) = (f+g)(v) + (f+g)(w), \\ (f+g)(\lambda v) &= f(\lambda v) + g(\lambda v) = \lambda f(v) + \lambda g(v) = \lambda(f(v) + g(v)) = \lambda(f+g)(v). \end{aligned}$$

b) follows by analogous computations.
As discussed in Example 6.7.(2), the space W^V of all maps from V to W is a vector space over K. By a) and b), $L(V,W)$ is a linear subspace of W^V, hence itself a vector space over K. □

Theorem 7.5. *Let U, V and W be vector spaces over K and let $f : U \to V$ and $g : V \to W$ be linear maps. Then $g \circ f : U \to W$ is a linear map.*

Proof. Let $v_1, v_2 \in U$. Then

$$\begin{aligned}(g \circ f)(v_1 + v_2) = g(f(v_1 + v_2)) &\overset{f \text{ linear}}{=} g(f(v_1) + f(v_2)) \overset{g \text{ linear}}{=} g(f(v_1)) + g(f(v_2)) \\ &= (g \circ f)(v_1) + (g \circ f)(v_2).\end{aligned}$$

For $\lambda \in K$ and $v \in U$ we further compute

$$(g \circ f)(\lambda v) = g(f(\lambda v)) \overset{f \text{ linear}}{=} g(\lambda f(v)) \overset{g \text{ linear}}{=} \lambda g(f(v)) = \lambda (g \circ f)(v).$$

Hence, $g \circ f$ is linear. □

Theorem 7.6. *Let V and W be vector spaces over K and let $f : V \to W$ be a linear map.*

a) If U is a linear subspace of V, then $f(U) = \{w \in W \mid \exists u \in U \text{ with } f(u) = w\}$ is a linear subspace of W.

b) If U' is a linear subspace of W, then $f^{-1}(U') = \{v \in V \mid f(v) \in U'\}$ is a linear subspace of V.

Proof. a) Let $w_1, w_2 \in f(U)$ and let $u_1, u_2 \in U$ with $f(u_1) = w_1$, $f(u_2) = w_2$. Since U is a linear subspace of V, it holds that $u_1 + u_2 \in U$. Then

$$w_1 + w_2 = f(u_1) + f(u_2) \overset{f \text{ linear}}{=} f(u_1 + u_2) \in f(U).$$

Moreover, for all $\lambda \in K$ it holds that $\lambda u_1 \in U$ and thus

$$\lambda w_1 = \lambda f(u_1) \overset{f \text{ linear}}{=} f(\lambda u_1) \in f(U).$$

Thus, $f(U)$ is a linear subspace of W.

b) Let $v_1, v_2 \in f^{-1}(U')$. Then $u_1, u_2 \in U'$, where $u_1 := f(v_1)$, $u_2 := f(v_2)$. Since U' is a linear subspace of U', $u_1 + u_2 \in U'$ and thus

$$f(v_1 + v_2) \overset{f \text{ linear}}{=} f(v_1) + (v_2) = u_1 + u_2 \in U' \quad \Rightarrow \quad v_1 + v_2 \in f^{-1}(U').$$

Moreover, for all $\lambda \in K$ it holds that $\lambda u_1 \in U'$ and thus

$$f(\lambda v_1) = \lambda f(v_1) = \lambda u_1 \in U' \quad \Rightarrow \quad \lambda v_1 \in f^{-1}(U').$$

Thus, $f^{-1}(U')$ is a linear subspace of V.

□

There are two interesting special cases of Theorem 7.6 that tell us a lot about a linear map.

Definition 7.7. Let V and W be vector spaces over K and let $f : V \to W$ be a linear map. The *kernel of f* (or *null space of f*) is given by

$$\ker f := \{v \in V \mid f(v) = 0\} = f^{-1}(\{0\}).$$

The *image of f* (or *range of f*) is given by

$$\operatorname{im} f := f(V) = \{w \in W \mid \exists v \in V \text{ with } f(v) = w\}.$$

Corollary 7.8. *Let V and W be vector spaces over K and let $f : V \to W$ be a linear map. Then* $\ker f$ *is a linear subspace of V and* $\operatorname{im} f$ *is a linear subspace of W.*

Proof. Since $\{0\}$ is a linear subspace of W, it follows from Theorem 7.6.b) that $\ker f = f^{-1}(\{0\})$ is a linear subspace of W.
Since V is a linear subspace of itself, it follows from Theorem 7.6.a) that $\operatorname{im} f = f(V)$ is a linear subspace of V. □

Example 7.9. (1) Let V and W be vector spaces over a field K and let $f : V \to W$, $f(v) = 0$ for all $v \in V$. Then f is linear with

$$\ker f = V \qquad \text{and} \qquad \operatorname{im} f = \{0\}.$$

(2) Let V be a vector space over a field K and consider $\mathrm{id}_V : V \to V$. Then

$$\ker \mathrm{id}_V = \{0\} \qquad \text{and} \qquad \operatorname{im} f = V.$$

(3) Let $v \in \mathbb{R}^3 \setminus \{0\}$. Then $f : \mathbb{R} \to \mathbb{R}^3$, $f(t) = tv$, is a linear map between real vector spaces. Obviously, $\ker f = \{0\}$, while $\operatorname{im} f = \operatorname{span}(\{v\})$ is the unique line in $\mathbb{R}^3$ that goes through v and the origin.

(4) Let $p : \mathbb{R}^3 \to \mathbb{R}^3$, $p(x, 0, 0)$. One checks without difficulties that p is linear. Then

$$\ker p = \{(0, y, z) \in \mathbb{R}^3 \mid (y, z) \in \mathbb{R}^3\} = \{0\} \times \mathbb{R}^2 \qquad \text{(the } yz\text{-plane in } \mathbb{R}^3\text{)},$$
$$\operatorname{im} p = \{(x, 0, 0) \in \mathbb{R}^3 \mid x \in \mathbb{R}\} = \mathbb{R} \times \{(0, 0)\} \qquad \text{(the } x\text{-axis in } \mathbb{R}^3\text{)}.$$

(5) Let $a_1, a_2, a_3 \in \mathbb{R}$ such that $a_i \neq 0$ for some $i \in \{1, 2, 3\}$. Then

$$f : \mathbb{R}^3 \to \mathbb{R}, \qquad f(x_1, x_2, x_3) = a_1x_1 + a_2x_2 + a_3x_3,$$

is a linear map. Then

$$\ker f = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid a_1x_1 + a_2x_2 + a_3x_3 = 0\}$$

is a linear subspace of $\mathbb{R}^3$. One observes that it is geometrically given as a plane in $\mathbb{R}^3$ containing the origin.

(6) Let I be an open interval and let again $D : C^1(I, \mathbb{R}) \to C^0(I, \mathbb{R})$, $D(f) = f'$. By Corollary 4.25.a),

$$\ker D = \{\text{constant functions } I \to \mathbb{R}\}.$$

For each $f \in C^0(I, \mathbb{R})$ the fundamental theorem of calculus implies that there exists $F \in C^1(I, \mathbb{R})$ with $D(F) = f$, hence

$$\operatorname{im} D = C^0(I, \mathbb{R}).$$

(7) Consider $\mathbb{R}^{\mathbb{N}}$, the space of real sequences as a real vector space, as in Example 6.7.(3). Consider the map

$$S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}, \qquad S((x_n)_{n\in\mathbb{N}}) = (x_{n+1})_{n\in\mathbb{N}}, \quad \text{i.e. } S(x_1, x_2, x_3, \dots) = (x_2, x_3, x_4, \dots),$$

which "cuts off" the first element of the sequence. One checks that S is linear[1], which is left to the reader as an exercise. Then

$$\ker S = \{(x, 0, 0, 0, \dots) \in \mathbb{R}^{\mathbb{N}} \mid x \in \mathbb{R}\}$$

and $\operatorname{im} S = \mathbb{R}^{\mathbb{N}}$, since for each $(x_1, x_2, x_3, \dots) \in \mathbb{R}^{\mathbb{N}}$, it holds that

$$S(0, x_1, x_2, x_3, \dots) = (x_1, x_2, x_3, x_4, \dots).$$

The next theorem shows a part of the significance of kernel and image.

Theorem 7.10. *Let V and W be vector spaces over K and let $f : V \to W$ be a linear map.*

a) f is surjective if and only if $\operatorname{im} f = W$.

b) f is injective if and only if $\ker f = \{0\}$.

c) f is bijective if and only if $\ker f = \{0\}$ *and* $\operatorname{im} f = W$.

Proof. a) This is obvious.

b) "$\Rightarrow$": If f is injective, then there exists at most one $v_0 \in V$ with $f(v_0) = 0$. But since f is linear, it must hold that $f(0) = 0$ by Remark 7.2.(1). Hence, $\ker f = \{0\}$.

"$\Leftarrow$": Let $v_1, v_2 \in V$ with $f(v_1) = f(v_2)$. Then $f(v_1) - f(v_2) = 0$ and since f is linear, it follows that $f(v_1 - v_2) = 0$. Since $\ker f = \{0\}$, it follows that $v_1 - v_2 = 0$, i.e. $v_1 = v_2$. This shows that f is injective.

c) This follows from combining parts a) and b).

□

Bijective linear maps are particularly important and will be given a special name.

Definition 7.11. Let V and W be vector spaces over K. A bijective linear map $f : V \to W$ is called an *isomorphism*. We call V and W *isomorphic* and write $V \cong W$ if there exists an isomorphism from V to W.

Theorem 7.12. *Let V and W be vector spaces over K and let $f : V \to W$ be an isomorphism. Then $f^{-1} : W \to V$ is an isomorphism as well.*

Proof. Since f is bijective, its inverse map $f^{-1} : W \to V$ is well-defined and bijective. It only needs to be shown that f^{-1} is a linear map as well. Let $w_1, w_2 \in W$. Since f is bijective, there are unique vectors $v_1, v_2 \in V$ with $f(v_1) = w_1$ and $f(v_2) = w_2$. Then

$$f^{-1}(w_1 + w_2) = f^{-1}(f(v_1) + f(v_2)) \overset{f \text{ linear}}{=} f^{-1}(f(v_1 + v_2)) = v_1 + v_2 = f^{-1}(w_1) + f^{-1}(w_2).$$

Moreover, for all $\lambda \in K$ we compute

$$f^{-1}(\lambda w_1) = f^{-1}(\lambda f(v_1)) \overset{f \text{ linear}}{=} f^{-1}(f(\lambda v_1)) = \lambda v_1 = \lambda f^{-1}(w_1).$$

Thus, f^{-1} is a linear map, which shows the claim. □

[1] S is also called a *shift operator*.

Corollary/Definition 7.13. *Let V be a vector space over K and let*

$$\mathrm{GL}(V) := \{f \in \mathrm{Hom}(V,V) \mid f \text{ is an isomorphism}\}.$$

Then $(\mathrm{GL}(V), \circ)$ is a group, called the general linear group of V.

Proof. Let $f, g \in \mathrm{GL}(V)$. Since a composition of bijective maps is bijective, $g \circ f$ is bijective. By Theorem 7.5, $g \circ f$ is linear, so $g \circ f \in \mathrm{GL}(V)$, so the composition of maps restricts to a well-defined map $\circ : \mathrm{GL}(V) \times \mathrm{GL}(V) \to \mathrm{GL}(V)$.
By Theorem 7.12, $f^{-1} \in \mathrm{GL}(V)$ for all $f \in \mathrm{GL}(V)$. By definition, $f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}_V$. Moreover, $\mathrm{id}_V \circ f = f = f \circ \mathrm{id}_V$ for all $f \in \mathrm{GL}(V)$. Since the composition of maps is obviously associative, this shows that $(\mathrm{GL}(V), \circ)$ is a group with neutral element id_V and where the inverse element of f is given by the inverse map f^{-1} for each $f \in \mathrm{GL}(V)$. □

We conclude this section by showing that every linear map can be restricted to an isomorphism if we restrict both domain and codomain accordingly.

Theorem 7.14. *Let V and W be vector spaces over K and let $f : V \to W$ be linear. If $U \subset V$ is a complement of* $\ker f$*, then the restriction*

$$f|_U : U \to \mathrm{im}\, f$$

is an isomorphism.

Proof. We need to show that $f|_U$ is injective and surjective. Let $w \in \mathrm{im}\, f$ and $v \in V$ with $f(v) = w$. Since

$$V = U \oplus \ker f,$$

Theorem 6.38 yields that there are unique $v_1 \in U$ and $v_2 \in \ker f$ with $v = v_1 + v_2$. By linearity of f,

$$f|_U(v_1) = f(v_1) = f(v_1) + f(v_2) = f(v_1 + v_2) = f(v) = w.$$

Hence, for each $w \in \mathrm{im}\, f|_U$, such that $f|_U$ is surjective.
If $u \in U$ satisfies $f|_U(u) = 0$, then $f(u) = 0$, such that $u \in U \cap \ker f = \{0\}$, i.e. $u = 0$. Hence, $\ker f|_U = \{0\}$ and Theorem 7.10.b) yields that f is injective. □

7.2 Linear maps and generating sets

Our main focus in the next chapters will lie on linear maps between finite-dimensional vector spaces. In order to understand their behaviour, we first want to investigate the interaction of linear maps with bases and dimensions of vector spaces.

Throughout this section, let K be a field.

We first take a look at arbitrary generating sets of vector spaces in interaction with linear maps.

Theorem 7.15. *Let V and W be vector spaces over K and let E be a generating set of V. Let $f : V \to W$ be a linear map.*

a) $f(E)$ is a generating set of $\mathrm{im}\, f$*, i.e.* $\mathrm{span}(f(E)) = \mathrm{im}\, f$.

b) f is uniquely determined by $f|_E$, i.e. if $g : V \to W$ is a linear map with $f(e) = g(e)$ for each $e \in E$, then $f = g$.

Proof. a) It follows by Remark 6.11.(4) that $\operatorname{span}(f(E)) \subset \operatorname{span}(f(V)) = f(V) = \operatorname{im} f$, so it only means to show the other inclusion. Let $v \in V$. Since $V = \operatorname{span}(E)$, there are $v_1, \dots, v_k \in E$, $\lambda_1, \dots, \lambda_k \in K$, such that $v = \sum_{i=1}^{k} \lambda_i v_i$. Since f is linear, we derive that

$$f(v) = f\Big(\sum_{i=1}^{k} \lambda_i v_i\Big) = \sum_{i=1}^{k} \lambda_i f(v_i) \in \operatorname{span}(f(E)),$$

hence $\operatorname{im} f = f(V) \subset \operatorname{span}(f(E))$.

b) Let $g : V \to W$ be linear with $f|_E = g|_E$ and let $v \in V$ be arbitrary. As in a), let $v_1, \dots, v_k \in E$ and $\lambda_1, \dots, \lambda_k \in K$ with $v = \sum_{i=1}^{k} \lambda_i v_i$. Since both f and g are linear maps and since $f(v_i) = g(v_i)$ for each $i \in \{1, 2, \dots, k\}$ by assumption, we derive

$$f(v) = \sum_{i=1}^{k} \lambda_i f(v_i) = \sum_{i=1}^{k} \lambda_i g(v_i) = g(v).$$

This shows that $f = g$.

□

In the case of bases, we can make an even stronger statement.

Theorem 7.16. *Let V and W be vector spaces over K and let B be a basis of V. For every map $F : B \to W$ there exists a unique linear map $f : V \to W$ with $f(b) = F(b)$ for each $b \in B$.*

Proof. Uniqueness: Since every basis is a generating set, the uniqueness of such a map follows from Theorem 7.15.b).
Existence: Let $v \in V$. Since B is a basis, there are unique $\lambda_b \in K$, $b \in B$, with $\lambda_B = 0$ for all but finitely many b, such that $v = \sum_{b \in B} \lambda_b b$. Then put

$$f(v) := \sum_{b \in B} \lambda_b F(b).$$

We obtain a well-defined map $f : V \to W$ and still need to prove its linearity. If $w = \sum_{b \in B} \mu_b b \in V$, then

$$\begin{aligned} f(v + w) &= f\Big(\sum_{b \in B} (\lambda_b + \mu_b) b\Big) \\ &= \sum_{b \in B} (\lambda_b + \mu_b) F(b) = \sum_{b \in B} \lambda_b F(b) + \sum_{b \in B} \mu_b F(b) \\ &= f\Big(\sum_{b \in B} \lambda_b b\Big) + f\Big(\sum_{b \in B} \mu_b b\Big) = f(v) + f(w). \end{aligned}$$

For $\alpha \in K$ we further compute

$$f(\alpha v) = f\Big(\sum_{b \in B} \alpha \lambda_b b\Big) = \sum_{b \in B} \alpha \lambda_b F(b) = \alpha \sum_{b \in B} \lambda_b F(b) = \alpha f(v).$$

Hence, f is linear. □

Remark 7.17. We can rephrase Theorem 7.16 in the following way: given a basis B of V the map

$$L(V, W) \to W^B, \qquad f \mapsto f|_B,$$

is a bijection. (Recall that $W^B = \{\text{maps } B \to W\}$.)

In other words, given a basis of its domain, a linear map is completely determined by its values on the basis. In the following results, we will express the injectivity and bijectivity of a linear map purely in terms of bases.

Theorem 7.18. *Let V and W be vector spaces over K and let B be a basis of V. Let $f : V \to W$ be linear. Then the following statements are equivalent:*

(i) f is injective.

(ii) $f(B)$ is a basis of $\operatorname{im} f$.

Proof. (i) $\Rightarrow$ (ii): By Theorem 7.15.a), $f(B)$ is a generating set of $\operatorname{im} f$, so it suffices to show that $f(B)$ is linearly independent. Let $b_1, \dots, b_k \in B$ and $\lambda_1, \dots, \lambda_k \in K$ such that $\sum_{i=1}^{k} \lambda_i f(b_i) = 0$. Since f is linear, this yields

$$f\Big(\sum_{i=1}^{k} \lambda_i b_i\Big) = 0 \quad \overset{\text{(i)}}{\Rightarrow} \quad \sum_{i=1}^{k} \lambda_i b_i = 0.$$

Since B is linearly independent, this yields $\lambda_1 = \cdots = \lambda_k = 0$. Thus, $f(B)$ is linearly independent.
(ii) $\Rightarrow$ (i): By Theorem 7.10.b), it suffices to show that $\ker f = \{0\}$. Let $v \in \ker f$. Since B is a basis of V, there are $b_1, \dots, b_k \in B$, $\lambda_1, \dots, \lambda_k \in K$, such that $v = \sum_{i=1}^{k} \lambda_i b_i$ and by linearity of f we obtain

$$0 = f(v) = \sum_{i=1}^{k} \lambda_i f(b_i) \quad \overset{\text{(ii)}}{\Rightarrow} \quad \lambda_1 = \cdots = \lambda_k = 0.$$

Hence, $\ker f = \{0\}$. □

Corollary 7.19. *Let V and W be vector spaces over K and let B be a basis of V. Let $f : V \to W$ be linear. Then f is an isomorphism if and only if $f(B)$ is a basis of W.*

Proof. If f is an isomorphism, it follows from Theorem 7.18 that $f(B)$ is a basis of $\operatorname{im} f = W$.
If $f(B)$ is a basis of W, then $W = \operatorname{span}(f(B)) = \operatorname{im} f$, hence f is surjective. By Theorem 7.18, f is also injective, hence an isomorphism. □

Theorem 7.20. *Let V and W be vector spaces over K and assume that V is finite-dimensional. Let $f : V \to W$ be an isomorphism. Then*

$$\dim V = \dim W.$$

In particular, W is finite-dimensional.

Proof. Let B be a basis of V. By Corollary 7.19, $f(B)$ is a basis of W. Since B is finite, $f(B)$ is finite, so W is finite-dimensional. Since f is bijective, it follows that

$$\dim V = |B| = |f(B)| = \dim W.$$

□

Remark 7.21. Consider K^n as a vector space over K for each $n \in \mathbb{N}$. Then K^n and K^m are isomorphic if and only if $m = n$. While "$\Leftarrow$" is trivial, "$\Rightarrow$" follows from Theorem 7.20

Theorem 7.22 (dimension formula for linear maps)**.** *Let V and W be vector spaces over K and assume that V is finite-dimensional. If $f : V \to W$ is a linear map, then*

$$\dim V = \dim(\ker f) + \dim(\operatorname{im} f).$$

Proof. Since V is finite-dimensional, $\ker f$ has a complement U by Theorem 6.39. Then by Corollary 6.41

$$\dim V = \dim U + \dim \ker f.$$

By Theorem 7.14, the restriction $f|_U : U \to \operatorname{im} f$ is an isomorphism, hence

$$\dim U = \dim \operatorname{im} f$$

by Theorem 7.20. Inserting this into the previous equation shows the claim. □

Corollary 7.23. *Let V and W be vector spaces over K and assume that V is finite-dimensional. Let $f : V \to W$ be a linear map.*

a) If f is injective, then $\dim V \leq \dim W$.

b) If f is surjective, then $\dim V \geq \dim W$.

c) If $\dim V = \dim W$ *and if f is injective or surjective, then f is an isomorphism.*

Proof. a) If f is injective, then the dimension formula yields $\dim(\operatorname{im} f) = \dim V$. Since $\operatorname{im} f$ is a linear subspace of W, $\dim W \geq \dim(\operatorname{im} f) = \dim V$.

b) If f is surjective, then the dimension formula yields $\dim V = \dim(\ker f) + \dim W \geq \dim W$.

c) If f is surjective and $\dim V = \dim W$, then $\dim(\ker f) = \{0\}$, hence $\ker f = \{0\}$, hence f is injective as well by Theorem 7.10.b). If f is injective and $\dim V = \dim W$, then $\dim(\operatorname{im} f) = \dim W$ by the dimension formula. But then by Corollary 6.33, every basis of $\operatorname{im} f$ is a basis of W, which shows that $\operatorname{im} f = W$, so f is surjective.

□

So far, we have observed in this section that any two isomorphic finite-dimensional vector spaces over K are necessarily of the same dimension. We want to study the converse: are any two finite-dimensional vector spaces over K isomorphic? The answer is indeed yes as we shall see below. The key observation for this statment is that any n-dimensional vector space over K is indeed isomorphic to K^n.

Definition 7.24. Let V be a finite-dimensional vector space over K. An *ordered basis of V* is a basis $B = \{b_1, b_2, \ldots, b_n\}$ of V together with a total order of its elements. We denote an ordered basis by $(b_1, b_2, \ldots, b_n)$, where the order of the tuple coincides with the total order of the basis.

In the following we will let $(e_1, e_2, \ldots, e_n)$ again denote *the canonical basis of K^n* as defined in Example 6.22.(2), seen as an ordered basis where the order is given by the index.

Theorem 7.25. *Let V be a finite-dimensional vector space over K and let $n = \dim V$. For each ordered basis $(b_1, b_2, \ldots, b_n)$ of V there exists a unique isomorphism $\phi_B : V \to K^n$, such that*

$$\phi_B(b_i) = e_i \qquad \forall i \in \{1, 2, \ldots, n\}.$$

Moreover, the map

$$F : \{\text{ordered bases of } V\} \to \{\text{isomorphisms } V \to K^n\}, \qquad F(B) = \phi_B,$$

is a bijection.

Proof. Given an ordered basis $B = (b_1, b_2, \dots, b_n)$, it follows from Theorem 7.16 that there exists a unique linear map $\phi_B : V \to K^n$ with $\phi(b_i) = e_i$ for each $i \in \{1, 2, \dots, n\}$. Since $\phi_B(B) = \{e_1, \dots, e_n\}$ is a basis of K^n, it follows from Corollary 7.19 that ϕ_B is an isomorphism.
By the uniqueness property of the ϕ_B, the map F is injective. Conversely, if $\phi : V \to K^n$ is an isomorphism, then $B_\phi := (\phi^{-1}(e_1), \phi^{-1}(e_2), \dots, \phi^{-1}(e_n))$ is an ordered basis of V by Corollary 7.19 and it holds by definition that $F(B_\phi) = \phi$. Hence, F is surjective. □

Remark 7.26. Note that given an ordered basis $(b_1, \dots, b_n)$ of B, one computes that the isomorphism $\phi_B : V \to K^n$ from Theorem 7.25 is explicitly given by

$$\phi_B\Big(\sum_{i=1}^{n} \lambda_i b_i\Big) = (\lambda_1, \lambda_2, \dots, \lambda_n) \qquad \forall \lambda_1, \lambda_2, \dots, \lambda_n \in K.$$

Corollary 7.27. *Let V and W be vector spaces over K and assume that V is finite-dimensional. Then $V \cong W$ if and only if $\dim V = \dim W$.*

Proof. "⇒" is Theorem 7.20.
"⇐": Put $n := \dim V = \dim W$. By Theorem 7.25, there are isomorphisms $\phi_V : V \to K^n$ and $\phi_W : W \to K^n$. By Theorem 7.12, $\phi_W^{-1} : K^n \to W$ is again an isomorphism, so $\phi_W^{-1} \circ \phi_V : V \to W$ is a composition of isomorphisms, hence an isomorphism itself. This shows that $V \cong W$. □

We conclude this section by observing that the vector space of linear maps between any two finite-dimensional vector spaces is again finite-dimensional.

Theorem 7.28. *Let V and W be finite-dimensional vector spaces over K. Then $L(V, W)$ is finite-dimensional with*

$$\dim L(V, W) = (\dim V) \cdot (\dim W).$$

Proof. (This proof is not discussed in the lecture videos.)

Put $m := \dim V$ and $n := \dim W$. Let $\{v_1, v_2, \dots, v_m\}$ be a basis of V and let $\{w_1, w_2, \dots, w_n\}$ be a basis of W. For $i \in \{1, 2, \dots, m\}$ and $j \in \{1, 2, \dots, n\}$ we let $f_{ij} : V \to W$ be the unique linear map with

$$f_{ij}(v_i) = w_j, \qquad f_{ij}(v_k) = 0 \quad \text{for } k \neq i.$$

(Such a unique map exists by Theorem 7.16.) Let

$$B := \left\{ f_{ij} \in L(V, W) \mid i \in \{1, 2, \dots, m\},\ j \in \{1, 2, \dots, n\} \right\}.$$

We claim that B is a basis of $L(V, W)$. It is easy to see that B is linearly independent, so we will only show that B is a generating set of $L(V, W)$.
We observe that by definition.

$$f_{ij}\Big(\sum_{i=1}^{m} \lambda_i v_i\Big) = \sum_{i=1}^{m} f_{ij}(\lambda_i v_i) = \lambda_i w_j = f(\lambda_i v_i). \tag{7.1}$$

Let $f : V \to W$ be an arbitrary linear map. Let $\mu_{ij} \in K$, where $i \in \{1, 2, dots, m\}$, $j \in \{1, 2, \dots, n\}$ be given such that

$$f(v_i) = \sum_{j=1}^{n} \mu_{ij} w_j \qquad \forall i \in \{1, 2, \dots, m\},\ j \in \{1, 2, \dots, n\}.$$

(Such numbers exist and are unique by Theorem 6.19 , since $\{w_1, \dots, w_n\}$ is a basis of W.)
Let $v = \sum_{i=1}^{m} \lambda_i v_i \in V$, where $\lambda_1, \dots, \lambda_m \in K$. Then by linearity of f,

$$\begin{aligned} f(v) &= \sum_{i=1}^{m} \lambda_i f(v_i) = \sum_{i=1}^{m} \lambda_i \sum_{j=1}^{n} \mu_{ij} w_j = \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \mu_{ij} w_j = \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \mu_{ij} f_{ij}(v_i) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} \lambda_i f_{ij}(v_i) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} f_{ij}(\lambda_i v_i) \overset{(7.1)}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} f_{ij}(v) \\ \Rightarrow \quad f &= \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} f_{ij}. \end{aligned}$$

Hence, $f \in \operatorname{span}(B)$, so $W = \operatorname{span}(B)$, which we wanted to show. So B is a basis of $L(V, W)$ and consequently

$$\dim L(V, W) = |B| = m \cdot n = (\dim V) \cdot (\dim W).$$

□

7.3 Linear equations

In this section we want to go back to something that some people might consider the core of mathematics: equations and their solutions. We will use the previous results to study equations involving linear maps and develop some general observations about their solutions. We will also go back to single-variable calculus and apply the results to so-called linear ordinary differential equations.

Definition 7.29. Let K be a field, let V and W be vector spaces over K and let $f : V \to W$ be a linear map. For $b \in W$ the equation

$$f(x) = b$$

is called a *linear equation*. The equation is called *homogeneous* if $b = 0$ and *inhomogeneous* otherwise. The set $\mathcal{L}_{f,b} := \{x \in V \mid f(x) = b\}$ is called the *solution set* of the equation.

Theorem 7.30. *Let K be a field, let V and W be vector spaces over K and let $f : V \to W$ be a linear map. If $b \in W$ and $x_0 \in V$ with $f(x_0) = b$, then*

$$\mathcal{L}_{f,b} = x_0 + \ker f := \{x_0 + v \in V \mid v \in \ker f\}.$$

Proof. "$\supset$": Let $v \in \ker f$. Then by linearity of f

$$f(x_0 + v) = f(x_0) + f(v) = b + 0 = b$$

and thus $x_0 + v \in \mathcal{L}_{f,b}$.
"$\subset$": Let $x_0 \in \mathcal{L}_{f,b}$. Then for each $v \in \mathcal{L}_{f,b}$ we obtain by linearity of f that

$$f(v - x_0) = f(v) - f(x_0) = b - b = 0,$$

so $v - x_0 \in \ker f$. Hence, $v = x_0 + (v - x_0) \in x_0 + \ker f$. □

Remark 7.31. Note that $\mathcal{L}_{f,0} = \ker f$ in Theorem 7.30. In particular, this means that the solution set of a homogeneous linear equation is a linear subspace of the domain of f.

In the following, we want to study two classes of linear equations in greater detail.

Linear ordinary differential equations with constant coefficients

Since we have already worked our way through almost two chapters of abstract linear algebra, we want to take a little detour in this subsection and consider an application of the previous results in single-variable calculus, which we are already familiar with.

Let I be an open interval and let $K = \mathbb{R}$ or $K = \mathbb{C}$. In analogy with Example 7.3.(4), for each $k \in \mathbb{N}$ the map $D : C^k(I,K) \to C^{k-1}(I,K)$, $D(u) = u'$, is a linear map. For each $n \in \mathbb{N}$ and $k \geq n$ we put

$$D^n := \underbrace{D \circ D \circ \cdots \circ D}_{n \text{ times}} : C^k(I,K) \to C^{k-n}(I,K)$$

By Theorem 7.5, each D^n is a linear map.

Definition 7.32. Let $a_0, a_1, \ldots, a_n \in K$ with $a_n \neq 0$ and consider the map

$$L : C^n(I,K) \to C^0(I,K), \quad L(u) = a_n D^n(u) + a_{n-1} D^{n-1}(u) + \cdots + a_1 D(u) + a_0 u,$$

which is linear by Theorem 7.4. A linear equation of the form

$$L(u) = b \quad \Leftrightarrow a_n u^{(n)}(t) + a_{n-1} u^{(n-1)}(t) + \cdots + a_n u'(t) + a_0 u(t) = b(t) \quad \forall t \in I, \tag{7.2}$$

where $b \in C^0(I,K)$, is called a *linear ordinary differential equation with constant coefficients* of *order* n.

By Theorem 7.30, the solution set of (7.2) is given by $u_0 + \ker L$, where $u_0 \in C^n(I,K)$ satisfies $L(u_0) = b$. In particular, this means that if $u_1, \ldots, u_k \in C^n(I,K)$ satisfy $L(u_i) = 0$ for each i, then

$$u_0 + \sum_{i=1}^{k} \lambda_i u_i$$

is a solution of $L(u) = b$ for all $\lambda_1, \ldots, \lambda_k \in K$. This observation is interpreted in physics as the *superposition principle*, which occurs for example in the consideration of wave interferences.

Example 7.33. (1) Given $g \in C^0(\mathbb{R},\mathbb{R})$, the equation

$$D(f) = g \quad \Leftrightarrow f'(t) = g(t) \quad \forall t \in I \tag{7.3}$$

is a linear ordinary differential equation (ODE) of first order. As we have seen in Corollary 4.25.a), $\mathcal{L}_{D,0} = \ker D$ is precisely the set of constant functions. By the fundamental theorem of calculus, the function $f_0 : I \to \mathbb{R}$, $f_0(t) = \int_0^t g(s)\, ds$, is a solution of (7.3). Using Theorem 7.30 it follows that

$$\mathcal{L}_{D,g} = \Big\{ \int_0^t g(s)\, ds + c \;\Big|\; c \in \mathbb{R} \Big\}.$$

(2) Let $c \in \mathbb{R}$ and $L : C^1(\mathbb{R},\mathbb{R}) \to C^0(\mathbb{R},\mathbb{R})$, $L(u) = u' - cu$. As we have seen in Corollary 4.25.c), the homogeneous linear differential equation of first order

$$L(u) = 0 \quad \Leftrightarrow u'(t) - cu(t) = 0 \quad \forall t \in \mathbb{R}$$

has the solution set

$$\mathcal{L}_{L,0} = \{ f_a \in C^1(\mathbb{R},\mathbb{R}) \mid a \in \mathbb{R} \}, \quad \text{where } f_a(t) = ae^{ct} \;\; \forall t \in \mathbb{R}.$$

From this description one easily checks explicitly the observation that we have made abstractly in this chapter that $\mathcal{L}_{L,0} = \ker L$ is a linear subspace of $C^1(\mathbb{R},\mathbb{R})$.

(3) Let $a_0, a_1 \in \mathbb{R}$ with $a_0 > 0$ and $a_1 \geq 0$ and consider

$$H : C^2(\mathbb{R}, \mathbb{R}) \to C^0(\mathbb{R}, \mathbb{R}), \qquad H(u) = u'' + a_1 u' + a_0 u.$$

Let $g \in C^0(\mathbb{R}, \mathbb{R})$ and consider the linear differential equation of second order

$$H(u) = g \quad \Leftrightarrow \quad u''(t) + a_1 u'(t) + a_0 u(t) = g(t) \quad \forall t \in \mathbb{R}$$

In the homogeneous case, i.e. for $g = 0$, this equation describes the motion of a damped harmonic oscillator, as you should have encountered in the lecture Theoretical Physics 1. In the inhomogenenous case, this equation describes a driven damped harmonic oscillator, where g encodes the effect of the externally applied force on the oscillator..

The following theorem shows that solutions of linear differential equations are uniquely determined by so-called initial conditons on their values and the values of their derivatives at one specific point.

Theorem 7.34 (Uniqueness theorem for linear ODEs)**.** *Let $u_1, u_2 : I \to K$ be solutions of* (7.3)*. If there exists $t_0 \in I$ with*

$$u_1(t_0) = u_2(t_0), \quad u_1'(t_0) = u_2'(t_0), \quad \dots, \quad u_1^{(n-1)}(t_0) = u_2^{(n-1)}(t_0),$$

then $u_1(t) = u_2(t)$ for all $t \in I$.

To prove this theorem, we need the following slightly technical lemma.

Lemma 7.35. *Let I be an interval and let $f : I \to \mathbb{R}$ be differentiable. If there exist $c > 0$ and $t_0 \in I$, such that*

$$|f'(t)| \leq c|f(t)| \quad \forall t \in I \quad \text{and} \quad f(t_0) = 0,$$

then $f(t) = 0$ for all $t \in I$.

Proof. We first consider the case that $f(t) \in [0, +\infty)$ for all $t \in I$, such that in particular $f'(t) \leq cf(t)$ for all $t \in I$. Then put $g : I \to \mathbb{R}$, $g(t) := f(t)e^{-ct}$. Then $g(t) \geq 0$ for all $t \in I$. Morever, g is differentiable and by the product rule

$$g'(t) = f'(t)e^{-ct} - cf(t)e^{-ct} = (f'(t) - cf(t))e^{-ct} \quad \forall t \in I.$$

By assumption, this yields $g'(t) \leq 0$ for all $t \in I$, hence g is monotonically decreasing by Corollary 4.22.b). Since $g(t_0) = f(t_0)e^{-ct_0} = 0$ and g is non-negative, this yields that $g(t) = 0$ for all $t \geq t_0$ and thus $f(t) = 0$ for all $t \geq t_0$.
Analogously, if we consider $\tilde{g} : I \to \mathbb{R}$, $\tilde{g}(t) = f(t)e^{ct}$. Since $-f'(t) \leq |f'(t)| \leq cf(t)$ for all $t \in I$, we derive analogously that

$$\tilde{g}'(t) = (f'(t) + cf(t))e^{ct} \geq 0,$$

so $\tilde{g}$ is monotonically increasing by Corollary 4.22.a). Since $\tilde{g}$ is non-negative and $\tilde{g}(t_0) = 0$, this shows that $\tilde{g}(t) = 0$ and thus $f(t) = 0$ for all $t \leq t_0$. Thus, in the case of non-negative f, we have shown that $f(t) = 0$ for all $t \in I$.
If $f : I \to \mathbb{R}$ takes negative values as well, put $h : I \to [0, +\infty)$, $h(t) := (f(t))^2$. Then $h(t_0) = 0$ and

$$|h'(t)| = |2f(t)f'(t)| = 2|f(t)| \cdot |f'(t)| \leq 2c|f(t)|^2 = 2c|h(t)|.$$

Applying the previous argument to h instead of f then shows $h(t) = 0$ and thus $f(t) = 0$ for all $t \in I$. □

Proof of Theorem 7.34. Put $u : I \to K$, $u(t) := u_1(t) - u_2(t)$ for all $t \in I$. By Theorem 7.30, u satisfies the homogeneous linear differential equation

$$a_n u^{(n)}(t) + a_{n-1}u^{(n-1)}(t) + \cdots + a_1 u'(t) + a_0 u(t) = 0 \quad \forall t \in I. \tag{7.4}$$

Put

$$f : I \to [0, +\infty), \qquad f(t) = |u(t)|^2 + |u'(t)|^2 + \cdots + |u^{(n-1)}(t)|^2 = \sum_{k=0}^{n-1} u^{(k)}(t)\overline{u^{(k)}(t)}.$$

By assumption on the derivatives of u_1 and u_2 it holds that $u^{(k)}(t_0) = 0$ for all $k \in \{0, 1, \ldots, n-1\}$, hence $f(t_0) = 0$. Moreover,

$$|u^{(k)}(t)| \leq \sqrt{f(t)} \quad \forall k \in \{0, 1, \ldots, n-1\},\ t \in I. \tag{7.5}$$

Thus, with $A := \max\{\frac{|a_0|}{|a_n|}, \frac{|a_1|}{|a_n|}, \ldots, \frac{|a_{n-1}|}{|a_n|}\} > 0$, it follows from (7.4) that for all $t \in I$,

$$|u^{(n)}(t)| \leq \frac{1}{|a_n|} \sum_{k=0}^{n-1} |a_k||u^{(k)}(t)| \leq A \sum_{k=0}^{n-1} |u^{(k)}(t)| \leq nA\sqrt{f(t)}. \tag{7.6}$$

Since conjugates of differentiable functions are differentiable, one sees that f is differentiable with

$$f'(t) = \sum_{k=0}^{n-1} \left(u^{(k+1)}(t)\overline{u^{(k)}(t)} + u^{(k)}(t)\overline{u^{(k+1)}(t)} \right)$$

Thus, by the triangle inequality of $|\cdot|$ and by the properties of absolute value, see Theorem 1.73,

$$\begin{aligned} |f'(t)| &\leq \sum_{k=0}^{n-2} 2|u^{(k+1)}(t)| \cdot |u^{(k)}(t)| + 2|u^{(n)}(t)| \cdot |u^{(n-1)}(t)| \\ &\overset{(7.5),(7.6)}{\leq} \sum_{k=0}^{n-2} 2\sqrt{f(t)} \cdot \sqrt{f(t)} + 2nA\sqrt{f(t)} \cdot \sqrt{f(t)} \\ &\leq 2(n-1)|f(t)| + 2nA|f(t)| = 2(n-1+nA)|f(t)|. \end{aligned}$$

Hence, if we put $C := 2(n-1+nA) > 0$, then $|f'(t)| \leq C|f(t)|$ for all $t \in I$. Since $f(t_0) = 0$, it follows from Lemma 7.35 that $f(t) = 0$ for all $t \in I$. This particularly yields that $u(t) = 0$ and thus $u_1(t) = u_2(t)$ for all $t \in I$. □

Corollary 7.36. *Let I be an open interval and $t_0 \in I$. Let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $L : C^n(I, K) \to C^0(I, K)$ be given as in* (7.2) *Then the map*

$$\phi : \ker L \to K^n, \qquad \phi(u) = (u(t_0), u'(t_0), \ldots, u^{(n-1)}(t_0)),$$

is linear and injective.

Proof. It follows straight from the linearity of derivatives that ϕ is linear. By Theorem 7.34, ϕ is injective. □

Corollary 7.37. *Let I be an open interval, $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $a_0, a_1, \ldots, a_n \in K$ with $a_n \neq 0$. The solution set of*

$$a_n u^{(n)}(t) + a_{n-1}u^{(n-1)}(t) + \cdots + a_0 u(t) = 0$$

is a linear subspace of $C^n(I, K)$ of dimension at most n.

Proof. Since the solution set is given by $\ker L$ in the situation of Corollary 7.36, it follows from that corollary and from Corollary 7.23 that $\dim \ker f \leq \dim K^n = n$. □

In fact, it holds in Corollary 7.36 that ϕ is an isomorphism and thus in Corollary 7.37 that the dimension of the solution set is precisely n. We will not prove these statements at this point, but (hopefully) derive them from a much more general statement towards the end of this course.

Linear systems / Systems of linear equations

Let K be a field, let $m, n \in \mathbb{N}$ and let $\{a_1, \ldots, a_n\} \subset K^m$. One checks (this is left to the reader as an exercise) that the map

$$A : K^n \to K^m, \quad A(x_1, \ldots, x_n) = x_1 \cdot a_1 + \cdots + x_n \cdot a_n,$$

is linear.
Given $b \in K^m$ we can then investigate linear equations of the form

$$A(x) = b.$$

Such an equation for maps between K^n and K^m can always be rephrased as a family of linear equations in K. For each $i \in \{1, 2, \ldots, m\}$ we let $a_{1i}, a_{2i}, \ldots, a_{mi} \in K$ denote the components of a_i, i.e. $a_i = (a_{1i}, a_{2i}, \ldots, a_{mi})$ and $b_1, \ldots, b_m \in K$ denote the components of b. Then, one computes that $A(x) = b$ if and only if all of the following equations hold:

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
\vdots \quad \vdots \qquad\qquad \vdots \quad \vdots & \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m.
\end{aligned}$$

Such a family is called a *linear system* or a *system of linear equations*. In the following chapters, we will learn a technique of how to solve such systems, the so-called Gaussian elimination or Gaussian algorithm that you might have already seen in high school. In the next chapter, we will introduce a convenient formalism to handle such linear systems and other questions regarding linear maps $K^n \to K^m$.

Chapter 8

Matrices

8.1 Matrices and linear maps $K^n \to K^m$

In this section we introduce the formalism of matrices. A matrix is first of all a smart way of arranging numbers in rectlangles, such that the numbers can be used to study linear maps as we shall see later on.

Definition 8.1. Let K be a field and $m, n \in \mathbb{N}$.

a) An $m \times n$-*matrix over* K is a family of numbers $a_{ij} \in K$, where $i \in \{1, 2, \ldots, m\}$, $j \in \{1, 2, \ldots, n\}$, which is organized as a rectangular array

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mn} \end{pmatrix} \tag{8.1}$$

We sometimes denote A by $(a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ or simply (a_{ij}).

b) We call the a_{ij} the *entries of* A, i the *row index* and j the *column index of* a_{ij}.

c) A $1 \times n$-matrix $\begin{pmatrix} b_1 & b_2 & \ldots & b_n \end{pmatrix}$ is called a *row vector*. With respect to A from (8.1), the row vectors $\begin{pmatrix} a_{i1} & a_{i2} & a_{i3} & \ldots & a_{in} \end{pmatrix}$, where $i \in \{1, 2, \ldots, m\}$, are called the *row vectors of* A or simply *rows of* A.

An $m \times 1$-matrix $\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$ is called a *column vector*. With respect to A from (8.1), the column vectors $\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$, where $j \in \{1, 2, \ldots, n\}$, are called the *column vectors of* A or simply *columns of* A.

d) The set of all $m \times n$-matrices over K is denoted by $M(m \times n, K)$.

e) An $m \times n$-matrix over $\mathbb{R}$ is called a *real $m \times n$-matrix*, while an $m \times n$-matrix over $\mathbb{C}$ is called a *complex $m \times n$-matrix*.

Example 8.2. $\begin{pmatrix} 1 & 2 & 3 \\ -1 & 3 & 0 \end{pmatrix}$ is a real 2×3-matrix whose rows are given by $\{(1 \quad 2 \quad 3), (-1 \quad 3 \quad 0)\}$ and whose columns are given by $\left\{\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \end{pmatrix}\right\}$.

$\begin{pmatrix} 1+i & 5 \\ 2i & -4 \end{pmatrix}$ is a complex 2×2-matrix, whose rows are $\{(1+i \quad 5), (2i \quad -4)\}$ and whose columns are $\left\{\begin{pmatrix} 1+i \\ 2i \end{pmatrix}, \begin{pmatrix} 5 \\ -4 \end{pmatrix}\right\}$.

Similar to elements of K^n, we can add matrices and multiply them by scalars. Not only that, the addition and scalar multiplication then form vector space structures on sets of matrices.

Theorem 8.3. *Let K be a field and $m, n \in \mathbb{N}$. Then $(M(m \times n, K), +, \cdot)$ is a vector space over K, where $+ : M(m \times n, K) \times M(m \times n, K) \to M(m \times n, K)$ and $\cdot : K \times M(m \times n, K) \to M(m \times n, K)$ are given by*

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+a_{22} & \dots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{pmatrix},$$

$$\lambda \cdot \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}.$$

The proof of this theorem is left to the reader as an exercise.

Remark 8.4. Let K be a field and $m, n \in \mathbb{N}$. One checks without difficulties that the maps

$$\Phi_1 : M(1 \times n, K) \to K^n, \quad \Phi_1 \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix} = (a_1, a_2, \dots, a_n),$$

$$\Phi_2 : M(m \times 1, K) \to K^m \quad \Phi_2 \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = (a_1, a_2, \dots, a_m),$$

are isomorphisms of vector spaces over K. In the following, we will thus occasionally identify K^n with $M(1 \times n, K)$ or $M(n \times 1, K)$ by implicitly using these isomorphisms.

The first observation showing the importance of matrices is the following theorem, which basically shows that matrices can be seen as linear maps $K^n \to K^m$ and vice versa.

Theorem 8.5. *Let K be a field and let $m, n \in \mathbb{N}$.*

a) Let $A \in M(m \times n, K)$ and let $(a_1, \dots, a_n)$ be the columns of A. Then

$$f_A : K^n \to K^m, \qquad f_A(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i \Phi_2(a_i),$$

is a linear map, where $\Phi_2 : M(m \times 1, K) \to K^m$ is given as in Remark 8.4.

b) Given a linear map $f : K^n \to K^m$, there exists a unique $A \in M(m \times n, K)$ with $f = f_A$.

Proof. a) This is shown by a straightforward computation and left to the reader as an exercise.

b) Let $(e_1, \dots, e_n)$ be the canonical basis of K^n. Since f is linear, it holds for all $(x_1, \dots, x_n) \in K^n$ that

$$f(x_1, \dots, x_n) = f\Big(\sum_{i=1}^{n} x_i e_i\Big) = \sum_{i=1}^{n} x_i f(e_i).$$

Thus, if we let A denote the matrix whose column vectors are $(\Phi_2^{-1}(f(e_1)), \dots, \Phi_2^{-1}(f(e_n)))$, then $f = f_A$ by definition of f_A. Moreover, A is uniquely determined by this condition. □

Remark 8.6. (1) Let $A \in M(m \times n, K)$ and let $\{e_1, \dots, e_n\}$ be the canonical basis of K^n. Up to identifying column vectors with elements of K^m, it follows from its construction that $f_A(e_j)$ *is precisely the j-th column vector of A for each* $j \in \{1, 2, \dots, n\}$.

(2) As shorthand notation, given $A \in M(m \times n, K)$ we will write

$$Av := f_A(v) \quad \forall v \in K^n,$$

where f_A is given as in Theorem 8.5.

Example 8.7. (1) Let $m, n \in \mathbb{N}$. The zero map $z : K^n \to K^m$, $z(x) = 0$ is induced by the zero matrix, i.e. the matrix for which every entry is 0.

(2) Let K be a field. The identity map $\mathrm{id}_{K^n} : K^n \to K^n$ satisfies $\mathrm{id}_{K^n} = f_{I_n}$, where

$$I_n \in M(n \times n, K), \qquad I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

I_n is called the *identity matrix of size n*.

(3) Let $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x, y) = (2x + y, x - 4y)$. Then, up to identifying column vectors with elements of $\mathbb{R}^2$,

$$f(x, y) = x\begin{pmatrix} 2 \\ 1 \end{pmatrix} + y\begin{pmatrix} 1 \\ -4 \end{pmatrix},$$

hence $f = f_A$, where $A = \begin{pmatrix} 2 & 1 \\ 1 & -4 \end{pmatrix}$.

Theorem 8.8. *Let $m, n \in \mathbb{N}$ and $A \in M(m \times n, K)$. Let $\{a_1, \dots, a_n\}$ denote the set of column vectors of A. Then $f_A : K^n \to K^m$ has the following properties:*

a) f_A is injective if and only if $\{a_1, \dots, a_n\}$ is linearly independent.

b) f_A is surjective if and only if $\{a_1, \dots, a_n\}$ is a generating set of $K^m (\cong M(m \times 1, K))$.

c) f_A is an isomorphism if and only if $m = n$ and $\{a_1, \dots, a_n\}$ is a basis of K^n.

Proof. Let $\{e_1, \dots, e_n\}$ be the canonical basis of K^n. By Remark 8.6, up to identifying columns with points in K^m, it holds that $\{a_1, \dots, a_n\} = \{f_A(e_1), \dots, f_A(e_n)\}$. We can thus trace the statements back to results from the previous chapter.

a) This follows from Theorem 7.18.

b) This follows from combining Theorem 7.15.a) and Theorem 7.10.a).

c) This follows from Corollary 7.19.

□

Theorem/Definition 8.9. *Let $\ell, m, n \in \mathbb{N}$ and let $A = (a_{ij}) \in M(\ell \times m, K)$ and $B = (b_{ij}) \in M(m \times n, K)$. Let $f_A : K^m \to K^\ell$ and $f_B : K^n \times K^m$ be the linear maps induced by A and B. Then*

$$f_A \circ f_B = f_C,$$

where $f_C : K^n \to K^\ell$ is the map induced by $C = (c_{ij}) \in M(\ell \times n, K)$, where

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj} \qquad \forall i \in \{1, 2, \ldots, \ell\},\ j \in \{1, 2, \ldots, n\}.$$

This matrix C is called the matrix product of A and B *and denoted by $C = A \cdot B$.*

Proof. First of all, by Theorem 7.5 $f_A \circ f_B : K^n \to K^\ell$ is a linear map, so by Theorem 8.5 there exists $C \in M(\ell \times n, K)$ with $f = f_C$, hence C is well-defined. If we denote the entries of C by $C = (c_{ij})$, then by definition

$$f_C(x_1, \ldots, x_n) = \sum_{j=1}^{n} x_j \begin{pmatrix} c_{1j} \\ c_{2j} \\ \vdots \\ c_{\ell j} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_j c_{1j} \\ \sum_{j=1}^{n} x_j c_{2j} \\ \vdots \\ \sum_{j=1}^{n} x_j c_{\ell j} \end{pmatrix}. \tag{8.2}$$

for all $(x_1, \ldots, x_n) \in K^n$. We further compute:

$$(f_A \circ f_B)(x_1, \ldots, x_n) = f_A\Big(\sum_{j=1}^{n} x_j b_{1j}, \sum_{j=1}^{n} x_j b_{2j}, \ldots, \sum_{j=1}^{n} x_j b_{mj} \Big) = \sum_{k=1}^{m} \Big(\sum_{j=1}^{n} x_j b_{kj} \Big) \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{\ell k} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^{m} (\sum_{j=1}^{n} x_j b_{kj}) a_{1k} \\ \sum_{k=1}^{m} (\sum_{j=1}^{n} x_j b_{kj}) a_{2k} \\ \vdots \\ \sum_{k=1}^{m} (\sum_{j=1}^{n} x_j b_{kj}) a_{\ell k} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} x_j \sum_{k=1}^{m} a_{1k} b_{kj} \\ \sum_{j=1}^{n} x_j \sum_{k=1}^{m} a_{2k} b_{kj} \\ \vdots \\ \sum_{j=1}^{n} x_j \sum_{k=1}^{m} a_{\ell k} b_{kj} \end{pmatrix}$$

Comparing this with (8.2) shows the claim. □

Remark 8.10. Let $\ell, m, n \in \mathbb{N}$, $A \in M(\ell \times m, K)$ and $B \in M(m \times n, K)$.

(1) Note that in the shorthand notation introduced in Remark 8.6.(2), the matrix product of satisfies

$$A(Bv) = (A \cdot B)v \qquad \forall v \in K^n.$$

One often writes simply $AB := A \cdot B$ for the matrix product.

(2) In order for the product of two matrices to be well-defined, *the number of columns of the first factor must coincide with the number of rows of the second factor*. In particular, this shows that while $A \cdot B$ is well-defined, $B \cdot A$ would only be well-defined if $\ell = n$.

(3) The definition of the coefficients of $A \cdot B$ already gives us a method of computing them. The (i, j)-entry of $A \cdot B$, i.e. the entry with row index i and column index j, is obtained by multiplying the entries of the i-th row of A one-by-one with the entries of the j-th column of B and add up all the product of the entries, as the following example will (hopefully) make clear.

Example 8.11. We consider the matrices $A \in M(3 \times 3, \mathbb{R})$ and $B \in M(3 \times 2, \mathbb{R})$ given by

$$A = \begin{pmatrix} 2 & 1 & -3 \\ 1 & 0 & 5 \\ 3 & -1 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 \\ 4 & 2 \\ 7 & -2 \end{pmatrix}.$$

Then $A \cdot B \in M(3 \times 2, \mathbb{R})$ is computed by the scheme of Remark 8.10.(3) in the following way.

$$\begin{array}{ccc} & 1 & -1 \\ & 4 & 2 \\ & 7 & -2 \\ 2 \quad 1 \quad -3 & \begin{pmatrix} 2\cdot 1 + 1\cdot 4 - 3\cdot 7 & 2\cdot(-1) + 1\cdot 2 + (-3)\cdot(-2) \\ 1\cdot 1 + 0\cdot 4 + 5\cdot 7 & 1\cdot(-1) + 0\cdot 2 + 5\cdot(-2) \\ 3\cdot 1 - 1\cdot 4 + 4\cdot 7 & 3\cdot(-1) - 1\cdot 2 + 4\cdot(-2) \end{pmatrix} & = \begin{pmatrix} -15 & 6 \\ 36 & -11 \\ 27 & -13 \end{pmatrix}. \\ 1 \quad 0 \quad 5 & & \\ 3 \quad -1 \quad 4 & & \end{array}$$

Hence, $A \cdot B = \begin{pmatrix} -15 & 6 \\ 36 & -11 \\ 27 & -13 \end{pmatrix}$.

Theorem 8.12. *Let $k, \ell, m, n \in \mathbb{N}$.*

a) Let $A \in M(k \times \ell, K)$, $B \in M(\ell \times m, K)$ and $C \in M(m \times n, K)$. Then

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \in M(k \times n, K).$$

b) Let $A \in M(m \times n, K)$. Then

$$I_m \cdot A = A \cdot I_n = A.$$

Proof. a) Since $A \cdot B$ induces $f_A \circ f_B$ and $B \cdot C$ induces $f_B \circ f_C$ by Theorem/Definition 8.9, it follows that $(A \cdot B) \cdot C$ induces $(f_A \circ f_B) \circ f_C$ while $A \cdot (B \cdot C)$ induces $f_A \circ (f_B \circ f_C)$. Since the composition of maps is associative,

$$(f_A \circ f_B) \circ f_C = f_A \circ (f_B \circ f_C),$$

which shows the claim.[1]

b) Since I_n induces id_{K^n} by Example 8.7.(2), it follows from Theorem 8.9 that $A \cdot I_n$ induces $f_A \circ \mathrm{id}_{K^n}$ while $I_m \cdot A$ induces $\mathrm{id}_{K^m} \circ f_A$. The claim thus follows, since

$$f_A \circ \mathrm{id}_{K^n} = \mathrm{id}_{K^m} \circ f_A = f_A.$$

□

[1] The reader is invited to perform an alternative proof of this part via a long, but straightforward computation using nothing but the definition of the matrix product.

For the rest of this section we shall study matrix representations of *isomorphisms*.

Definition 8.13. Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$. Such an A is called a *quadratic matrix*. We call A *invertible* if $f_A : K^n \to K^n$ is an isomorphism. If $B \in M(n \times n, K)$ is the matrix with $f_A^{-1} = f_B$, then we call B *the inverse matrix of* A and put $A^{-1} := B$. The set

$$\mathrm{GL}(n, K) := \{A \in M(n \times n, K) \mid A \text{ is invertible}\}$$

is called the *general linear group of degree n over K*.

Remark 8.14. Let K be a field and $n \in \mathbb{N}$. Then by Theorem 8.8.c) a matrix $A \in M(n \times n, K)$ is invertible if and only if the column vectors of A form a basis of K^n.

Theorem 8.15. *Let K be a field, let $n \in N$.*

a) Let $A \in \mathrm{GL}(n, K)$. Then $A \cdot A^{-1} = A^{-1} \cdot A = I_n$ and $(A^{-1})^{-1} = A$.

b) Let $A, B \in \mathrm{GL}(n, K)$. Then $A \cdot B \in \mathrm{GL}(n, K)$ and $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$.

Proof. a) By Theorem/Definition 8.9, $A \cdot A^{-1}$ is the matrix inducing $f_A \circ f_A^{-1}$, while $A^{-1} \cdot A$ induces $f_A^{-1} \circ f_A$. Since

$$f_A \circ f_A^{-1} = f_A^{-1} \circ f_A = \mathrm{id}_{K^n},$$

the first claim follows from Example 8.7.(2). The second claim follows from the evident statement $(f_A^{-1})^{-1} = f_A$.

b) Since $f_A, f_B : K^n \to K^n$ are isomorphisms, $f_A \circ f_B$ is an isomorphism as well. It is induced by $A \cdot B$, hence $(A \cdot B)^{-1}$ induces $(f_A \circ f_B)^{-1} = f_B^{-1} \circ f_A^{-1}$. By definition, $f_B^{-1} \circ f_A^{-1}$ is induced by $B^{-1} \cdot A^{-1}$, which shows the claim.

□

Corollary 8.16. $(\mathrm{GL}(n, K), \cdot)$ *is a group for each $n \in \mathbb{N}$*

Proof. By Theorem 8.15.b), the matrix product restricts to an operation

$$\cdot : \mathrm{GL}(n, K) \times \mathrm{GL}(n, K) \to \mathrm{GL}(n, K).$$

The associativity of $\cdot$ and the existence of a neutral element follow from the two parts of Theorem 8.12, while the existence of inverses is a consequence of Theorem 8.15.a). □

There are indeed methods of explicitly computing the inverse of a given invertible matrix. These methods require some additional preparations, but we will learn about them later in this course.

8.2 Matrix representations of linear maps

Throughout this section, we let K be a field and V and W be finite-dimensional vector spaces over K and put $n = \dim V$ and $m = \dim W$.

We recall from Theorem 7.25 that each ordered basis B of V defines an isomorphism from V to K^n, which displays the coefficients of a vector with respect to the basis B as coordinates in K^n.

Definition 8.17. Given an ordered basis $B = (b_1, \dots, b_n)$ of V, we call the isomorphism

$$\phi_B : V \to K^n, \qquad \phi_B\Big(\sum_{j=1}^{n} \lambda_j b_j\Big) = (\lambda_1, \dots, \lambda_n),$$

the *coordinate representation of* B.

If B is an ordered basis of V and C be an ordered basis of W and if $\phi_B : V \to K^n$ and $\phi_C : W \to K^m$ are their coordinate representations, then for each $f \in L(V, W)$, the map

$$\phi_C \circ f \circ \phi_B^{-1} : K^n \to K^m$$

is a composition of linear maps, hence linear. By Theorem 8.5, it is thus induced by a matrix.

Definition 8.18. Let B be an ordered basis of V, let C be an ordered basis of W and let $f \in L(V, W)$. The unique matrix $A \in M(m \times n, K)$ with

$$f_A = \phi_C \circ f \circ \phi_B^{-1}$$

is called *the matrix representation of f with respect to B and C*. We will also denote it by

$$M_{B,C}(f) := A.$$

Remark 8.19. A common way of writing the relation of maps in the previous definition is in the form of a so-called *commutative diagram* as

$$\begin{array}{ccc} V & \xrightarrow{\ f\ } & W \\ {\scriptstyle \phi_B}\downarrow & & \downarrow{\scriptstyle \phi_C} \\ K^n & \xrightarrow{\ M_{B,C}(f)\ } & K^m \end{array}$$

Commutativity here means that following the arrows from the top-left corner to the bottom-right corner in both possible ways gives the same result, i.e. $\phi_C \circ f = M_{B,C}(f) \circ \phi_B$ (where we identify $M_{B,C}(f)$ with the linear map $K^n \to K^m$ that it induces).

Theorem 8.20. *Let $B = (b_1, \dots, b_n)$ be an ordered basis of V and $C = (c_1, \dots, c_m)$ be an ordered basis of W and let $f \in L(V, W)$. If $M_{B,C}(f) = (a_{ij}) \in M(m \times n, K)$, then*

$$f(b_j) = \sum_{i=1}^{m} a_{ij} c_i \qquad \forall j \in \{1, 2, \dots, n\}.$$

Proof. By Remark 8.6.(1), the j-th column of $M_{B,C}(f)$ satisfies

$$(a_{1j}, a_{2j}, \dots, a_{mj}) = (\phi_C \circ f \circ \phi_B^{-1})(e_j) = \phi_C(f(b_j)) \qquad \forall j \in \{1, 2, \dots, n\}.$$

Thus, by Remark 7.26,

$$f(b_j) = \phi_C^{-1}(a_{1j}, a_{2j}, \dots, a_{mj}) = \sum_{i=1}^{m} a_{ij} c_i.$$

□

Example 8.21. We will now use Theorem 8.20 to compute our first example of matrix representations. Assume that $\dim V = 2$. Let $\lambda \in K$ and let $f \in L(V,V)$, $f(v) = \lambda v$ for all $v \in V$. Let $B = (b_1, b_2)$ be a basis of V and define another basis $C = (c_1, c_2)$ of V by $c_1 := -b_1$ and $c_2 := 2b_2$. Then

$$f(b_1) = \lambda b_1 = -\lambda \cdot c_1 + 0 \cdot c_2, \qquad f(b_2) = \lambda b_2 = 0 \cdot c_1 + \frac{\lambda}{2} \cdot c_2.$$

Thus,

$$M_{B,C}(f) = \begin{pmatrix} -\lambda & 0 \\ 0 & \frac{\lambda}{2} \end{pmatrix}.$$

If we define another basis $D = (d_1, d_2)$ of V by $d_1 = b_1$, $d_2 = b_1 + b_2$, then

$$f(b_1) = \lambda b_1 = \lambda \cdot d_1 + 0 \cdot d_2, \qquad f(b_2) = \lambda b_2 = -\lambda b_1 + \lambda(b_1 + b_2) = -\lambda \cdot d_1 + \lambda \cdot d_2.$$

Hence,

$$M_{B,D}(f) = \begin{pmatrix} \lambda & -\lambda \\ 0 & \lambda \end{pmatrix}.$$

Since $c_1 = -d_1$, $c_2 = 2(d_2 - d_1)$, we further obtain from

$$f(c_1) = \lambda c_1 = -\lambda d_1 = -\lambda \cdot d_1 + 0 \cdot d_2, \qquad f(c_2) = \lambda c_2 = 2\lambda(d_2 - d_1) = -2\lambda \cdot d_1 + 2\lambda \cdot d_2,$$

that

$$M_{C,D}(f) = \begin{pmatrix} -\lambda & -2\lambda \\ 0 & 2\lambda \end{pmatrix}.$$

We next want to investigate a special case of matrix representations, which abstractly describes the *change of coordinates* in a vector space. Roughly speaking, such a transformation matrix, as we shall call it, contains the information requires to take vectors given in terms of one basis and translate them into vectors given in terms of another basis.

Definition 8.22. Let $B = (b_1, \ldots, b_n)$ and $B' = (b'_1, \ldots, b'_n)$ be ordered bases of V. The matrix representation

$$T_{B,B'} := M_{B,B'}(\mathrm{id}_{K^n}) \in GL(n, K),$$

i.e. the matrix inducing the map $\phi_{B'} \circ \mathrm{id}_{K^n} \circ \phi_B^{-1} = \phi_{B'} \circ \phi_B : K^n \to K^n$, is called the *transformation matrix from B to B'*.

Corollary 8.23. *Let $B = (b_1, \ldots, b_n)$ and $B' = (b'_1, \ldots, b'_n)$ be ordered bases of V. If $T_{B,B'} = (t_{ij}) \in \mathrm{GL}(n, K)$ is the transformation matrix from B to B'. Then*

$$b_j = \sum_{i=1}^{n} t_{ij} b'_i \qquad \forall j \in \{1, 2, \ldots, n\}.$$

Proof. This follows directly from Theorem 8.20 with $f = \mathrm{id}_{K^n}$. □

Remark 8.24. By definition of transformation matrices, it holds for all ordered bases B and B' of V that

$$T_{B',B} = T_{B,B'}^{-1}.$$

Example 8.25. Let $V = K^3$, let $B = (e_1, e_2, e_3)$ the canonical basis of K^3 and let

$$B' = ((1,0,0),(1,1,0),(1,1,1)).$$

One checks that B' is linearly independent with $|B'| = 3 = \dim K^3$, hence B' is a basis of K^3. Let $T_{B,B'} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \in M(3 \times 3, K)$ be the transformation matrix. Then, by Corollary 8.23,

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = e_1 = t_{11} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + t_{21} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + t_{31} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Evidently, the unique solution of this equation is $t_{11} = 1$, $t_{21} = 0$ and $t_{31} = 0$. Moreover,

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = e_2 = t_{12} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + t_{22} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + t_{32} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

which yields $t_{32} = 0$, $t_{22} = 1$ and $t_{12} = -1$. Analogously,

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = e_3 = t_{13} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + t_{23} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + t_{33} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

which is uniquely solved by $t_{33} = 1$, $t_{23} = -1$ and $t_{13} = 1$. Thus,

$$T_{B,B'} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

In the same way, one computes that $T_{B',B} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, which is left to the reader as an exercise.

We have seen how linear maps can be represented as matrices with respect to choices of coordinates and we have seen how changes of coordinates can be describes in terms of matrices. This leads to a question combining these two constructions: How are the matrix representations of different bases, i.e. of different choices of coordinates, related to each other?

Theorem 8.26. *Let B and B' be ordered bases of V and C and C' be ordered bases of W. Let $f \in L(V,W)$ and let $M_{B,C}(f), M_{B',C'}(f) \in M(m \times n, K)$ be the corresponding matrix representations. Then*

$$M_{B',C'}(f) = T_{C,C'} \cdot M_{B,C}(f) \cdot T_{B,B'}^{-1}.$$

Proof. $M_{B',C'}(f)$ is the matrix which induces $\phi_{C'} \circ f \circ \phi_{B'}^{-1} \in L(K^n, K^m)$. We compute:

$$\begin{aligned} \phi_{C'} \circ f \circ \phi_{B'}^{-1} &= \phi_{C'} \circ \mathrm{id}_W \circ f \circ \mathrm{id}_V \circ \phi_{B'}^{-1} \\ &= \phi_{C'} \circ \phi_C^{-1} \circ \phi_C \circ f \circ \phi_B^{-1} \circ \phi_B \circ \phi_{B'}^{-1} \\ &= (\phi_{C'} \circ \mathrm{id}_W \circ \phi_C^{-1}) \circ (\phi_C \circ f \circ \phi_B^{-1}) \circ (\phi_B \circ \mathrm{id}_V \circ \phi_{B'}^{-1}). \end{aligned}$$

By Theorem/Definition 8.9, the latter composition is induced by

$$M_{C,C'}(\mathrm{id}_W) \cdot M_{B,C}(f) \cdot M_{B',B}(\mathrm{id}_V) = T_{C,C'} \cdot M_{B,C}(f) \cdot T_{B',B} = T_{C,C'} \cdot M_{B,C}(f) \cdot T_{B,B'}^{-1}.$$

□

It turns out that as a generalization of the constructions of the previous section, matrix representations of compositions are given as products of matrix representations.

Theorem 8.27. *Let X be another finite-dimensional vector space over K and let $\ell = \dim X$. Let B be an ordered basis of V, let C be an ordered basis of W and let D be an ordered basis of X. Then for all $f \in L(V,W)$ and $g \in L(W,X)$ it holds that*

$$M_{B,D}(g \circ f) = M_{C,D}(g) \cdot M_{B,C}(f) \in M(\ell \times n, K).$$

Proof. By definition, $M_{B,D}(g \circ f)$ is the matrix inducing the map

$$\phi_D \circ g \circ f \circ \phi_B^{-1} = \phi_D \circ g \circ \mathrm{id}_V \circ f \circ \phi_B^{-1} = (\phi_D \circ g \circ \phi_C^{-1}) \circ (\phi_C \circ f \circ \phi_{B^{-1}}).$$

By Theorem/Definition 8.9 and the definition of the two matrices, the latter composition is induced by $M_{C,D}(f) \cdot M_{B,C}(f)$, so the claim follows from the uniqueness of the matrices inducing linear maps. □

Our next aim is to show that each linear map $V \to W$ can be represented by a matrix in a particularly simple way for convenient choice of bases of V and W. For this purpose, we need to introduce some additional terminology.

Definition 8.28. a) Let $f : V \to W$ be linear. The *rank of f* is given by

$$\operatorname{rank} f := \dim(\operatorname{im} f).$$

b) Let $A \in M(m \times n, K)$ and let $(a_1, \ldots, a_n)$ denote its column vectors. The *rank of A* is given by

$$\operatorname{rank} A := \dim(\operatorname{span}(\{a_1, \ldots, a_n\})).$$

Remark 8.29. (1) If $A \in M(m \times n, K)$ and $f_A : K^n \to K^m$ is the induced linear map, then

$$\operatorname{rank} A = \operatorname{rank} f_A.$$

This follows straight from the definitions.

(2) Let again $A \in M(m \times n, K)$ and let $(a_1, \ldots, a_n)$ denote its column vectors. Then $\{a_1, \ldots, a_n\}$ is a generating set of $\operatorname{im} f$, so by Corollary 6.27.c) there exists $B \subset \{a_1, \ldots, a_n\}$ which is a basis of $\operatorname{im} f$, i.e. such that B is linearly independent. Then $\operatorname{rank} A = |B|$, so $\operatorname{rank} A$ is the *maximal number of linearly independent column vectors of A*.

(3) The rank of $f \in L(V,W)$ satisfies

$$\operatorname{rank} f \leq \min\{\dim V, \dim W\}.$$

Evidently, $\operatorname{rank} f \leq \dim W$, since $\operatorname{im} f$ is a linear subspace of W. If B is a basis of V, then by Theorem 7.15.a), $f(B)$ is a generating set of $\operatorname{im} f$, hence $\operatorname{rank} f \leq |f(B)| \leq |B| = \dim V$.

We want to generalize the equation $\operatorname{rank} A = f_A$ by showing that the rank of a linear map $V \to W$ coincides with the rank of each of its matrix representations.

Theorem 8.30. *Let V' and W' be finite-dimensional vector spaces over K, let $\psi : V \to V'$ and $\phi : W \to W'$ be isomorphisms and let $f \in L(V, W)$. Then*

$$\operatorname{rank}(\phi \circ f \circ \psi) = \operatorname{rank} f.$$

Proof. Since ϕ is an isomorphism, it restricts to an isomorphism $\operatorname{im}(f \circ \psi) \xrightarrow{\cong} \operatorname{im}(\phi \circ f \circ \psi)$, hence $\operatorname{rank}(f \circ \psi) = \operatorname{rank}(\phi \circ f \circ \psi)$. Since ψ is an isomorphism, $\ker(f \circ \psi) = \ker f$, so by the dimension formula for linear maps

$$\operatorname{rank}(f \circ \psi) = \dim V - \dim(\ker(f \circ \psi)) = \dim V - \dim(\ker f) = \operatorname{rank} f.$$

□

Corollary 8.31. *Let B be a basis of V, let C be a basis of W and let $\phi_B : V \to K^n$ and $\phi_C : W \to K^m$ be their coordinate representations. Let $f \in L(V, W)$. Then*

$$\operatorname{rank} f = \operatorname{rank}(\phi_C \circ f \circ \phi_B^{-1}) = \operatorname{rank} M_{B,C}(f).$$

Proof. This follows from combining Theorem 8.30 with Remark 8.29.(1). □

In particular, the rank of $M_{B,C}(f)$ is independent of the choices of bases B and C.

Definition 8.32. Let $k, \ell, m, n \in \mathbb{N}$. Let $A = (a_{ij}) \in M(k \times m, K)$, $B = (b_{ij}) \in M(k \times n, K)$, $C = (c_{ij}) \in M(\ell \times m, K)$ and $D = (d_{ij}) \in M(\ell \times n, K)$. We define a matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix} \in M((k+\ell) \times (m+n), K)$ by

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & b_{11} & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2m} & b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{km} & b_{k1} & b_{k2} & \dots & b_{kn} \\ c_{11} & c_{12} & \dots & c_{1m} & d_{11} & d_{12} & \dots & d_{1n} \\ c_{21} & c_{22} & \dots & c_{2m} & d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{\ell 1} & c_{\ell 2} & \dots & c_{\ell m} & d_{\ell 1} & d_{\ell 2} & \dots & d_{\ell n} \end{pmatrix}$$

i.e. by putting all the entries of A, B, C and D into one big matrix according to the scheme indicated by the notation $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. A matrix given in this form is called a *block matrix*.

In the following theorem we let $0_{k \times \ell} \in M(k \times \ell, K)$ denote the zero matrix with k rows and ℓ entries for all $k, \ell \in \mathbb{N}$.

Theorem 8.33. *Let $f : V \to W$ be a linear map and let $r := \operatorname{rank} f$. Then there are bases B of V and C of W, such that*

$$M_{B,C}(f) = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I_r & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} \longleftarrow r\text{-th row}$$

Proof. Since V is finite-dimensional, there exists a complement U of $\ker f$ and by Theorem 7.14, f restricts to an isomorphism $f|_U : U \to \operatorname{im} f$. Let $(c_1, \dots, c_r)$ be a basis of $\operatorname{im} f$. Then $(b_1, \dots, b_r)$ is a basis of U, where $b_j = (f|_U)^{-1}(c_j)$ for each j. By the dimension formula for linear maps, $\dim(\ker f) = n - r$ and we let $(b_{r+1}, \dots, b_n)$ be a basis of $\ker f$. Then $B := (b_1, \dots, b_r, b_{r+1}, \dots, b_n)$ is a basis of V. By the basis extension theorem, we can extend $(c_1, \dots, c_r)$ to a basis $C := (c_1, \dots, c_r, c_{r+1}, \dots, c_m)$ of W.
Let $M_{B,C}(f) = (a_{ij})$. Then by Theorem 8.20 it holds for all $j \in \{1, 2, \dots, r\}$ that

$$c_j = f(b_j) = \sum_{i=1}^{m} a_{ij} c_i,$$

which shows that

$$a_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad \forall i \in \{1, 2, \dots, m\},\ j \in \{1, 2, \dots, r\}.$$

If $j > i$, then again by Theorem 8.20,

$$0 = f(b_j) = \sum_{i=1}^{m} a_{ij} c_i.$$

Since C is linearly independent, this yields $a_{ij} = 0$ whenever $j > r$. Together, these computations show the claim. □

Remark 8.34. In the situation of Theorem 8.33 the special form of the matrix representation implies that

$$(\phi_C \circ f \circ \phi_B^{-1})(x_1, x_2, \dots, x_n) = (x_1, x_2 \dots, x_r, 0, 0, \dots, 0).$$

The observation from Remark 8.34 has some interesting consequences in certain special cases.

Corollary 8.35. *Let $f \in L(V, W)$.*

a) If f is injective, then there will be ordered bases B of V and C of W, such that

$$(\phi_C \circ f \circ \phi_B^{-1})(x_1, \dots, x_n) = (x_1, \dots, x_n, 0, \dots, 0) \qquad \forall (x_1, \dots, x_n) \in K^n.$$

b) If f is surjective, then there will be ordered bases B of V and C of W, such that

$$(\phi_C \circ f \circ \phi_B^{-1})(x_1, \dots, x_n) = (x_1, \dots, x_m) \qquad \forall (x_1, \dots, x_n) \in K^n.$$

c) If f is an isomorphism, then there will be ordered bases B of V and C of W, such that

$$\phi_C \circ f \circ \phi_B^{-1} = \mathrm{id}_{K^n}.$$

Proof. a) If f is injective, then $n \leq m$ by Corollary 7.23.a) and $\operatorname{rank} f = n$ by the dimension formula. By Theorem 8.33, there are B and C, such that

$$M_{B,C}(f) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

One computes that $\phi_C \circ f \circ \phi_B^{-1}$ then takes the desired form.

b) If f is surjective, then $n \geq m$ by Corollary 7.23.b) and $\operatorname{rank} f = m$. By Theorem 8.33, there are B and C, such that

$$M_{B,C}(f) = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

and the claim follows in analogy with a).

c) If f is an isomorphism, then $n = m$ by Corollary 7.23.c) and $\operatorname{rank} f = n$. By Theorem 8.33, there are B and C, such that $M_{B,C}(f) = I_n$, which shows the claim.

□

To conclude, we shall study some consequences of Theorem 8.33 for matrices.

Corollary 8.36. *Let $m, n \in \mathbb{N}$, $A \in M(m \times n, K)$ and put $r := \operatorname{rank} A$. Then there are $S \in \mathrm{GL}(m, K)$ and $T \in \mathrm{GL}(n, K)$ with*

$$SAT^{-1} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. Consider the linear map $f_A : K^n \to K^m$. By Theorem 8.33 there are bases B of K^n and C of K^m, such that $\phi_C \circ f_A \circ \phi_B^{-1}$ is induced by

$$\begin{pmatrix} I_r & 0 \\ 0 & 0. \end{pmatrix},$$

where $\phi_B : K^n \to K^n$ and $\phi_C : K^m \to K^m$ are the coordinate representations. These are again induced by matrices. If $T \in \mathrm{GL}(n, K)$ induces ϕ_B and $S \in \mathrm{GL}(m, k)$ induces ϕ_C, then, by the previous results of this chapter, $\phi_C \circ f_A \circ \phi_B^{-1}$ is induced by $S \cdot A \cdot T^{-1}$. By uniqueness of the matrix inducing a linear map $K^n \to K^m$, this shows the claim. □

Corollary 8.37. *Let $m, n \in \mathbb{N}$ and $A, B \in M(m \times n, K)$. Then $\operatorname{rank} A = \operatorname{rank} B$ if and only if there are $S \in \mathrm{GL}(m, K)$ and $T \in \mathrm{GL}(n, K)$ with*

$$SAT^{-1} = B.$$

Proof. "⇐": This follows by applying Theorem 8.30 to the induced maps.
"⇒": Put $r := \operatorname{rank} A = \operatorname{rank} B$. By Corollary 8.36, there are matrices $S_1, S_2 \in \mathrm{GL}(m, K)$ and $T_1, T_2 \in \mathrm{GL}(n, K)$, such that

$$S_1 A T_1^{-1} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = S_2 B T_2^{-1} \quad \Rightarrow \quad S_2^{-1} S_1 A T_1^{-1} T_2 = B.$$

So the claim follows with $S := S_1 S_2^{-1} \in \mathrm{GL}(m, K)$ and $T := T_2^{-1} T_1 \in \mathrm{GL}(n, K)$, since by the properties of Theorem 8.15,

$$T^{-1} = (T_2^{-1} T_1)^{-1} = T_1^{-1} (T_2^{-1})^{-1} = T_1^{-1} T_2.$$

□

8.3 Dual spaces and transposes of matrices

Throughout this section, let K be a field.

We have already seen that sets of linear maps between vector spaces are themselves vector spaces. In this section, we want to focus on a particular case of spaces of linear maps and see some results about matrix representations for these spaces.

Definition 8.38. Let V be a vector space over K. We call

$$V^* := L(V, K) = \mathrm{Hom}(V, K)$$

the dual space of V. Elements of V^* are called *linear forms*.

Remark 8.39. By Theorem 7.4 applied to $W = K$, the dual space V^* is a vector space over K with respect to pointwise addition and scalar multiplication. If V is finite-dimensional, then V^* is finite-dimensional with

$$\dim V = \dim V^*$$

by Theorem 7.28.

Example 8.40. Let $\varphi \in (K^n)^* = L(K^n, K)$. By linearity,

$$\varphi(x_1, \ldots, x_n) = \varphi\Big(\sum_{i=1}^{n} x_i e_i\Big) = \sum_{i=1}^{n} x_i \varphi(e_i) \quad \forall (x_1, \ldots, x_n) \in K^n.$$

Conversely, by Theorem 7.16 for all $a_1, \ldots, a_n \in K$ there exists a unique $\varphi \in (K^n)^*$ with $\varphi(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i$ for all $(x_1, \ldots, x_n) \in K^n$. Consequently,

$$(K^n)^* = \Big\{ \varphi : K^n \to K \;\Big|\; \exists a_1, \ldots, a_n \in K \text{ with } \varphi(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i \;\; \forall (x_1, \ldots, x_n) \in K^n \Big\}.$$

Theorem/Definition 8.41. *Let V be a finite-dimensional vector space over K and let $(b_1, \ldots, b_n)$ be a basis of V. Then $(b_1^*, \ldots, b_n^*)$ is a basis of V^*, where for each $i \in \{1, 2, \ldots, n\}$ we let $b_i^* : V \to K$ be the unique linear map with*

$$b_i^*(b_j) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad \forall j \in \{1, 2, \ldots, n\}.$$

$(b_1^*, \ldots, b_n^*)$ is called the dual basis to $(b_1, \ldots, b_n)$*. The symbol δ_{ij}, defined as above, is sometimes called the* Kronecker delta.

Proof. (This proof is not discussed in the lecture videos.)
One checks that the dual basis is a special case of the basis constructed in the proof of Theorem 7.28, so the claim follows as a special case of that theorem. □

Example 8.42. Consider the ordered basis $B = (b_1, b_2)$ of $\mathbb{R}^2$ as a real vector space, where $b_1 = (2, 1)$, $b_2 = (3, 1)$. We want to find the dual basis $B^* = (\varphi_1, \varphi_2)$ to B of $(\mathbb{R}^2)^*$, where $\varphi_1 := b_1^*$ and $\varphi_2 := b_2^*$. To write φ_1 and φ_2 in a form described in Example 8.40, we recall that

$$\varphi(x_1, x_2) = x_1 \varphi(1, 0) + x_2 \varphi(0, 1) \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

Since $(1,0) = b_2 - b_1$ and $(0,1) = 3b_1 - 2b_2$, we compute from the linearity of φ_1 and φ_2 that

$$\begin{aligned}
\varphi_1(1,0) &= \varphi_1(b_2) - \varphi_1(b_1) = 0 - 1 = -1,\\
\varphi_1(0,1) &= 3\varphi_1(b_1) - 2\varphi_1(b_2) = 3 \cdot 1 - 2 \cdot 0 = 3,\\
\varphi_2(1,0) &= \varphi_2(b_2) - \varphi_2(b_1) = 1 - 0 = 1,\\
\varphi_2(0,1) &= 3\varphi_2(b_1) - 2\varphi_2(b_2) = 3 \cdot 0 - 2 \cdots 1 = -2.
\end{aligned}$$

Hence, $\varphi_1, \varphi_2 : \mathbb{R}^2 \to \mathbb{R}$ are given by

$$\varphi_1(x_1, x_2) = -x_1 + 3x_2, \quad \varphi_2(x_1, x_2) = x_1 - 2x_2 \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

The following result explores the connection between dual bases and matrix representations of linear maps.

Theorem 8.43. *Let V and W be finite-dimensional vector spaces over K and let $f \in L(V,W)$. If $B = (b_1, \ldots, b_n)$ is a basis of V and $C = (c_1, \ldots, c_m)$ is a basis of W, then $M_{B,C}(f) \in M(m \times n, K)$ is given as $M_{B,C}(f) = (a_{ij})$, where*

$$a_{ij} = c_i^*(f(b_j)) \qquad \forall i \in \{1, 2, \ldots, m\},\ j \in \{1, 2, \ldots, n\}.$$

Proof. Let $M_{B,C}(f) = (a_{ij}) \in M(m \times n, K)$. By Theorem 8.20,

$$f(b_j) = \sum_{k=1}^{m} a_{kj} c_k \quad \forall j \in \{1, 2, \ldots, n\}.$$

Thus, by linearity of the c_i^*, we obtain for all $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$ that

$$c_i^*(f(b_j)) = c_i^*\Big(\sum_{k=1}^{m} a_{kj} c_k\Big) = \sum_{k=1}^{m} a_{kj} c_i^*(c_k) = \sum_{k=1}^{m} a_{kj} \delta_{kj} = a_{ij}.$$

□

A crucial property of dual spaces is that any linear map between vector spaces induces a linear map between the dual spaces of these spaces, but goint "in the other direction".

Definition 8.44. Let V and W be vector spaces over K and let $f \in L(V,W)$. The map

$$f^t : W^* \to V^*, \qquad f^t(\varphi) = \varphi \circ f,$$

is called *the transpose of f*. (Note that f^t is well-defined, since the composition of linear maps is again linear.)

Theorem 8.45. *Let V and W be vector spaces over K and let $f : V \to W$ be a linear map.*

a) The transpose $f^t : W^* \to V^*$ is linear.

b) Let X be another vector space over K and let $g \in L(W,X)$. Then $(g \circ f)^t = f^t \circ g^t : X^* \to V^*$.

Proof. a) Let $\varphi_1, \varphi_2 \in W^* = L(W,K)$. Then for all $v \in V$:

$$(f^t(\varphi_1 + \varphi_2))(v) = (\varphi_1 + \varphi_2)(f(v)) = \varphi_1(f(v)) + \varphi_2(f(v)) = (f^t(\varphi_1))(v) + (f^t(\varphi_2))(v).$$

Thus $f^t(\varphi_1 + \varphi_2) = f^t(\varphi_1) + f^t(\varphi_2)$. Moreover, for all $\lambda \in K$ and $v \in V$ we compute

$$(f^t(\lambda\varphi_1))(v) = \lambda\varphi_1(f(v)) = \lambda(f^t(\varphi_1))(v),$$

which implies $f^t(\lambda\varphi_1) = \lambda f^t(\varphi_1)$. Thus, f^t is linear.

b) We compute for all $\varphi \in X^*$ that

$$(g \circ f)^t(\varphi) = \varphi \circ g \circ f = f^t(\varphi \circ g) = f^t(g^t(\varphi)) = (f^t \circ g^t)(\varphi).$$

□

Let $f \in L(V, W)$, let B be a basis of V, let C be a basis of W and let B^* and C^* denote the corresponding dual bases of V^* and W^*, respectively. Then $f : V \to W$ has a matrix representation

$$M_{B,C}(f) \in M(m \times n, K)$$

with respect to B and C. Since $\dim V^* = \dim V = n$ and $\dim W^* = \dim W = m$, the transpose $f^t : W^* \times V^*$ has a matrix representation

$$M_{C^*,B^*}(f^t) \in M(n \times m, K)$$

with respect to C^* and B^*. In the following, we want to investigate the relation between $M_{B,C}(f)$ and $M_{C^*,B^*}(f^t)$.

Definition 8.46. Let $m, n \in \mathbb{N}$ and $A = (a_{ij}) \in M(m \times n, K)$. The *transpose of A* is the matrix $A^t = (a'_{ij}) \in M(n \times m, K)$ given by

$$a'_{ij} = a_{ji} \qquad \forall i \in \{1, 2, \ldots, n\},\ j \in \{1, 2, \ldots, m\}.$$

Example 8.47. (1) If $A = \begin{pmatrix} 2 & 1 & 3i \\ 4 & 7i & 5 \end{pmatrix} \in M(2 \times 3, \mathbb{C})$, then $A^t = \begin{pmatrix} 2 & 4 \\ 1 & 7i \\ 3i & 5 \end{pmatrix} \in M(3 \times 2, \mathbb{C})$.

(2) For each $n \in \mathbb{N}$ it holds that $I_n^t = I_n$.

Theorem 8.48. *Let V and W be finite-dimensional vector spaces over K. Let B be an ordered basis of V, C be an ordered basis of W and let B^* and C^* denote the corresponding dual bases of V^* and W^*, respectively. Then*

$$M_{C^*,B^*}(f^t) = (M_{B,C}(f))^t.$$

Proof. Let $A := M_{B,C}(f) \in M(m \times n, K)$, $A = (a_{ij})$ and $\tilde{A} := M_{C^*,B^*}(f^t) \in M(n \times m, K)$, $\tilde{A} = (\tilde{a}_{ij})$. Then by Theorem 8.20,

$$f^t(c_i^*) = \sum_{k=1}^{n} \tilde{a}_{ki} b_k^* \quad \forall i \in \{1, 2, \ldots, m\}.$$

By Theorem 8.43, this yields for all $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$.

$$a_{ij} = c_i^*(f(b_j)) = (c_i^* \circ f)(b_j) = f^t(c_i^*)(b_j) = \sum_{k=1}^{n} \tilde{a}_{ki} b_k^*(b_j) = \sum_{k=1}^{n} a_{ki}\delta_{kj} = \tilde{a}_{ji}.$$

Consequently, $\tilde{A} = A^t$, which we wanted to show. □

Theorem 8.49. *Let $\ell, m, n \in \mathbb{N}$.*

a) For all $A \in M(\ell \times m, K)$ and $B \in M(m \times n, K)$ it holds that

$$(A \cdot B)^t = B^t \cdot A^t.$$

b) For all $A \in \mathrm{GL}(n,K)$ it holds that $A^t \in \mathrm{GL}(n,K)$ with $(A^t)^{-1} = (A^{-1})^t$.

c) For all $A \in M(m \times n, K)$ it holds that

$$\operatorname{rank} A = \operatorname{rank} A^t.$$

Proof. a) Let $A = (a_{ij})$, $B = (b_{ij})$ and put $C := A \cdot B = (c_{ij})$. Put $A^t = (a'_{ij})$, $B^t = (b'_{ij})$ and $C^t = (c'_{ij})$. By definition, $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$, hence

$$c'_{ij} = c_{ji} = \sum_{k=1}^{m} a_{jk} b_{ki} \qquad \forall i \in \{1,2,\dots,n\},\ j \in \{1,2,\dots,\ell\}.$$

We further compute that $B^t \cdot A^t = (d_{ij})$ is given by

$$d_{ij} = \sum_{k=1}^{m} b'_{ik} a'_{kj} = \sum_{k=1}^{m} b_{ki} a_{jk} = \sum_{k=1}^{m} a_{jk} b_{ki} = c'_{ij}$$

for all $i \in \{1,2,\dots,n\}$ and $j \in \{1,2,\dots,\ell\}$. Hence, $C^t = B^t \cdot A^t$.

b) By a), it holds that

$$A^t \cdot (A^{-1})^t = (A^{-1}A)t = I_n^t = I_n.$$

Analogously, $(A^{-1})^t \cdot A^t = (A \cdot A^{-1})^t = I_n$, hence $(A^{-1})^t = (A^t)^{-1}$.

c) Let $r := \operatorname{rank} A$. By Corollary 8.36, there are $S \in \mathrm{GL}(m,K)$ and $T \in \mathrm{GL}(n,K)$ with $SAT^{-1} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} =: C$. Since $C = C^t$, this shows that

$$C = (SAT^{-1})^t = (T^{-1})^t A^t S^t$$

by part a). By Corollary 8.37, this shows that $\operatorname{rank} A^t = \operatorname{rank} C = r$.

□

We conclude this section with a statement that might not look particularly exciting, but that we will make use of in the following section.

Corollary 8.50. *Let $m, n \in \mathbb{N}$ and $A \in M(m \times n, K)$ and let $(a_1^r, a_2^r, \dots, a_m^r)$ denote the row vectors of A. Then*

$$\operatorname{rank} A = \dim(\operatorname{span}(\{a_1^r, a_2^r \dots, a_m^r\})),$$

i.e. rank *A is the maximal number of linearly independent row vectors of A.*

Proof. By definition of A^t, up to identifying row vectors with column vectors, $(a_1^r, a_2^r, \dots, a_m^r)$ is given by the set of column vectors of A^t. Hence, the claim follows from Theorem 8.49. □

8.4 Gaussian elimination and systems of linear equations

Throughout this section, we let K be a field.

In this section, we will study a method which helps us solve various computational problems occurring in linear algebra, for example:

- Find the solution set of a system of linear equations.
- Compute the rank of a matrix.
- Let $A \in \mathrm{GL}(n,K)$. Compute A^{-1} explicitly.

Moreover, many questions of linear algebra like the linear independence of a finite set of vectors in K^n, constructions of dual bases etc. can be formulated as a study of solutions of systems of linear equations, as you will see in the exercise sheets.

8.4.1 Computation of the rank of a matrix

We begin by introducing some basic operations one can perform to the rows of a matrix.

Definition 8.51. Let $m, n \in \mathbb{N}$ and let $A \in M(m \times n, K)$.

a) For all $i, j \in \{1, 2, \ldots, m\}$ with $i \leq j$ we let $S_{i,j} : M(m \times n, K) \to M(m \times n, K)$ be the following map: if $A \in M(m \times n, K)$ has row vectors $(a_1^r, \ldots, a_{i-1}^r, a_i^r, a_{i+1}^r, \ldots, a_{j-1}^r, a_j^r, a_{j+1}^r, \ldots, a_n^r)$, then $S_{i,j}(A)$ has the row vectors

$$(a_1^r, \ldots, a_{i-1}^r, a_j^r, a_{i+1}^r, \ldots, a_{j-1}^r, a_i^r, a_{j+1}^r, \ldots, a_m^r)$$

i.e. $S_{i,j}(A)$ is obtained from A by swapping the i-th and the j-th row.

b) For all $i \in \{1, 2, \ldots, m\}$ and $\lambda \in K \setminus \{0\}$ we let $M_i^\lambda : M(m \times n, K) \to M(m \times n, K)$ be the following map: if $A \in M(m \times n, K)$ has row vectors $(a_1^r, \ldots, a_{i-1}^r, a_i^r, a_{i+1}^r, \ldots, a_n^r)$, then $M_i^\lambda(A)$ has row vectors

$$(a_1^r, \ldots, a_{i-1}^r, \lambda \cdot a_i^r, a_{i+1}^r, \ldots, a_m^r),$$

i.e. $M_{i,\lambda}(A)$ is obtained from A by multiplying its i-th row by λ.

c) For all $i, j \in \{1, 2, \ldots, m\}$ and $\lambda \in K \setminus \{0\}$, let $E_{i,j}^\lambda : M(m \times n, K) \to M(m \times n, K)$ be given as follows: if $A \in M(m \times n, K)$ has row vectors $(a_1^r, \ldots, a_{i-1}^r, a_i^r, a_{i+1}^r, \ldots, a_n^r)$, then $E_{i,j}^\lambda$ has row vectors

$$(a_1^r, \ldots, a_{i-1}^r, a_i^r + \lambda \cdot a_j^r, a_{i+1}^r, \ldots, a_m^r),$$

i.e. $E_{i,j}^\lambda(A)$ is obtained from A by adding the λ-fold multiple of the j-th row of A to its i-th row.

A map of one of one of the three forms given in a)-c) is called an *elementary row operation*.

Let $A, B \in M(m \times n, K)$. We say that *$B$ is obtained from A by elementary row operations* if there exists $k \in \mathbb{N}$ and elementary row operations $T_1, \ldots, T_k$, such that

$$B = (T_1 \circ T_2 \circ \cdots \circ T_k)(A).$$

In this case we also call A and B *row-equivalent*.

Example 8.52. Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \in M(3 \times 3, \mathbb{R})$. Then

$$S_{1,3}(A) = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad M_2^{-3}(A) = \begin{pmatrix} 1 & 1 & 0 \\ -3 & 0 & -3 \\ 0 & 1 & 1 \end{pmatrix}, \qquad E_{3,2}^2(A) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{pmatrix}.$$

The following result shows the relevance of elementary row operations for rank computations.

Theorem 8.53. *Let $A, B \in M(m \times n, K)$. Let $(a_1^r, \dots, a_m^r)$ be the row vectors of A and $(b_1^r, \dots, b_m^r)$ be the row vectors of B. If A and B are row-equivalent, then*

$$\operatorname{span}(\{a_1^r, \dots, a_m^r\}) = \operatorname{span}(\{b_1^r, \dots, b_m^r\}).$$

In particular,

$$\operatorname{rank} A = \operatorname{rank} B.$$

Proof. (This proof is not discussed in the lecture videos.)

Once the equality of the spans has been established, the equality of the ranks follows directly from Corollary 8.50, so we only need to show the first equation.
To do so, it suffices to show the claim in the case that B is obtained from A via one elementary row operation.

- If there are $i \leq j$, such that $B = S_{i,j}(A)$, then by construction

  $$\{b_1^r, \dots, b_m^r\} = \{a_1^r, \dots, a_{i-1}^r, a_j^r, a_{i+1}^r, \dots, a_{j-1}^r, a_i^r, a_{j+1}^r, \dots, a_m^r\} = \{a_1^r, \dots, a_m^r\},$$

  which obviously implies the claim.

- If there are $i \in \{1, 2, \dots, m\}$ and $\lambda \in K \setminus \{0\}$, such that $B = M_i^\lambda(A)$, then

  $$\{b_1^r, b_2^r, \dots, b_m^r\} = \{a_1^r, a_2^r, \dots, a_{i-1}^r, \lambda a_i^r, a_{i+1}^r, \dots, a_m^r\}$$

  Every linear combination $\sum_{j=1}^m \mu_j a_j^r$ can be written as

  $$\sum_{j=1}^m \mu_j a_j^r = \sum_{j=1}^{i-1} \mu_j b_j^r + \frac{\mu_i}{\lambda} b_i^r + \sum_{j=1}^{i-1} \mu_j b_j^r \in \operatorname{span}(\{b_1^r, \dots, b_m^r\}),$$

  hence $\operatorname{span}(\{a_1^r, \dots, a_m^r\}) \subset \operatorname{span}(\{b_1^r, \dots, b_m^r\})$. The other inclusion is shown completely analogously, which proves the claim.

- If there are $i, j \in \{1, 2, \dots, m\}$ and $\lambda \in K$ with $B = E_{i,j}^\lambda(A)$, then $b_k^r = a_k^r$ for all $k \neq i$, while $b_i^r = a_i^r + \lambda a_j^r$. Then for every linear combination $\sum_{k=1}^m \mu_k a_k^r$, $\mu_1, \dots, \mu_m \in K$, one computes that

  $$\sum_{k=1}^m \mu_k a_k^r = \sum_{k=1}^{j-1} \mu_k b_k^r + (\mu_j - \lambda \mu_i) b_j^r + \sum_{k=j+1}^m \mu_k b_k^r \in \operatorname{span}(\{b_1^r, m \dots, b_m^r\}).$$

  Hence $\operatorname{span}(\{a_1^r, \dots, a_m^r\}) \subset \operatorname{span}(\{b_1^r, \dots, b_m^r\})$. The other inclusion is shown completely analogously, which proves the claim.

Since in the general case, B is obtained from A by applying finitely many elementary row operations, one can successively apply the corresponding items out of this list and prove the claim in the general case. □

This theorem gives as an idea to compute the rank of a matrix A: apply a number of elementary row operations to A to obtain a matrix whose rank is "easy to determine".

Definition 8.54. Let $m, n \in \mathbb{N}$. A matrix $A = (a_{ij}) \in M(m \times n, K)$ is *in row-echelon form* if it has the following properties:

- there exists $r \in \{1, 2, \dots, m\}$, such that $a_{ij} = 0$ whenever $j > r$. (i.e. all entries below the r-th row of A are vanishing.
- for each $i \in \{1, 2, \dots, r\}$ there exists $j_i \in \{1, 2, \dots, n\}$ with $a_{ij} = 0$ for all $j < j_i$ and $a_{ij_i} \neq 0$ for all $i \in \{1, 2, \dots, r\}$, and it holds that

$$j_1 < j_2 < \cdots < j_r.$$

Explicitly, A is of the form

$$A = \begin{pmatrix} 0 & \dots & 0 & a_{1j_1} & \dots & a_{1(j_2-1)} & a_{1j_2} & \dots & a_{1(j_r-1)} & a_{1j_r} & \dots & a_{1n} \\ 0 & \dots & 0 & 0 & \dots & 0 & a_{2j_2} & \dots & a_{2(j_r-1)} & a_{2j_r} & \dots & a_{2n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & a_{rj_r} & \dots & a_{rn} \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}$$

The numbers a_{ij_i}, where $i \in \{1, 2, \dots, r\}$, are called *pivots* or *leading coefficients of* A.

Example 8.55. The following matrices are in row-echelon form:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 0 & 7 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 0 & 0 & 6 & 7 & 8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 9 \end{pmatrix}.$$

This one is not: $\begin{pmatrix} 1 & 0 & 2 \\ 0 & 3 & 4 \\ 5 & 6 & 0 \end{pmatrix}$.

Remark 8.56. If $A \in M(m \times n, K)$ is in row-echelon form and r is given as in Definition 8.54, then $\operatorname{rank} A = r$, since the properties of the pivots imply that the first r rows, which are the only non-vanishing ones, must be linearly independent.

The key observation is now contained in the following theorem.

Theorem 8.57 (Gaussian elimination)**.** *Let $m, n \in \mathbb{N}$ and $A \in M(m \times n, K)$. There is a matrix $B \in M(m \times n, K)$ in row-echelon form, such that B can be obtained from A by elementary row operations.*

Proof. (The method of this proof is called ***Gaussian algorithm*** *and is crucial to solving explicit problems in linear algebra.)*

Put $j_1 := \min\{j \in \{1, 2, \dots, n\} \mid a_{ij} \neq 0 \text{ for some } i \in \{1, 2, \dots, m\}\}$, i.e. the column index of the first column of A that has a non-zero entry. Let $i \in \{1, 2, \dots, m\}$ with $a_{ij_1} \neq 0$ and consider $A' = (a'_{ij}) := S_{1,i}(A)$. For each $i \in \{2, 3, \dots, m\}$ put $\lambda_i := -\frac{a'_{ij_1}}{a'_{1j_1}}$ and add the λ_i-fold multiple of the first row of A' to the i-th row of A'. Call the resulting matrix $A'' := E_{2,1}^{\lambda_2}(E_{3,1}^{\lambda_3}(\dots E_{m,1}^{\lambda_m}(A')\dots))$. Then A'' is of the form

$$A'' = \begin{pmatrix} 0 & \dots & 0 & b_{1j_1} & b_{1(j_1+1)} & \dots & b_{1n} \\ 0 & \dots & 0 & 0 & b_{2(j_1+1)} & \dots & b_{2n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & b_{m(j_1+1)} & \dots & b_{mn} \end{pmatrix}.$$

Put $A_1 := \begin{pmatrix} b_{2(j_1+1)} & \dots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{m(j_1+1)} & \dots & b_{mn} \end{pmatrix}$. If we apply the same method to A_1 that we applied to A, we obtain a matrix of the form

$$\begin{pmatrix} 0 & \dots & 0 & c_{1j_1} & \dots & c_{1(j_2-1)} & c_{1j_2} & c_{1(j_2+1)} & \dots & c_{1n} \\ 0 & \dots & 0 & 0 & \dots & 0 & c_{2j_2} & c_{2(j_2+1)} & \dots & c_{2n} \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & c_{3(j_2+1)} & \dots & c_{3n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & & \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & c_{m(j_2+1)} & \dots & c_{mn} \end{pmatrix}.$$

Put $A_2 := \begin{pmatrix} c_{3(j_2+1)} & \dots & c_{3n} \\ \vdots & \ddots & \vdots \\ c_{m(j_2+1)} & \dots & c_{mn} \end{pmatrix}$. We can iterate this procedure and apply the same method to A_2 and go on and on and on. After finitely many iteration steps, we ultimately obtain a matrix in row-echelon form that is obtained from A via elementary row operations. □

Remark 8.58. Note that in the situation of Theorem 8.57 the matrix B is *not* unique. For example, if B is in row-echelon form, then $M_i^\lambda(B)$ is in row-echelon form as well for all $i \in \{1, 2, \dots, m\}$ and $\lambda \in K \setminus \{0\}$.

We can combine Theorems 8.53 and Theorem 8.57 to get an explicit method of computing the rank of a matrix.

Method to compute the rank of $A \in M(m \times n, K)$:

- Use the Gaussian algorithm to transform A by elementary row-operations into a matrix B in row-echelon form.
- The rank of A is the index of the lowest row of B which contains a non-zero entry.

Example 8.59. Consider the following matrix$A \in M(4 \times 5, \mathbb{R})$:

$$A = \begin{pmatrix} 0 & 0 & 0 & -4 & -3 \\ 0 & -4 & 2 & 2 & 3 \\ 0 & 6 & -3 & 1 & 1 \\ 0 & 2 & -1 & 1 & 0 \end{pmatrix} \xrightarrow{S_{1,4}} \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 0 & -4 & 2 & 2 & 3 \\ 0 & 6 & -3 & 1 & 1 \\ 0 & 0 & 0 & -4 & -3 \end{pmatrix} \xrightarrow{E_{2,1}^2} \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 4 & 3 \\ 0 & 6 & -3 & 1 & 1 \\ 0 & 0 & 0 & -4 & -3 \end{pmatrix}$$

$$\xrightarrow{E_{3,1}^{-3}} \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 4 & 3 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & -4 & -3 \end{pmatrix} \xrightarrow{E_{4,2}^1} \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 4 & 3 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{E_{3,2}^{\frac{1}{2}}} \begin{pmatrix} 0 & 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 4 & 3 \\ 0 & 0 & 0 & 0 & \frac{5}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since this last matrix is in row-echelon form and has rank three, we derive that

$$\operatorname{rank} A = 3.$$

8.4.2 Computation of the inverse of an invertible matrix

Recall that every invertible matrix is necessarily a *quadratic* matrix, so we will mostly focus on these in this subsection. Before doing so, we want to give an alternative approach to elementary row operations. It turns out that every elementary row operation of a matrix A is given by the matrix product of A with a matrix of a particular form. The proof of the following result will be omitted, but can be given by purely elementary methods and explicitly computing the matrix products that occur.

Theorem/Definition 8.60. *Let $m, n \in \mathbb{N}$ and $A \in M(m \times n, K)$ and let again $I_m \in GL(m, K)$ denote the identity matrix of size m. If T is an elementary row operation, then*

$$T(A) = T(I_m) \cdot A.$$

A matrix of the form $T(I_m)$, where T is an elementary row operation, is called an elementary matrix.

For the sake of completeness, we write down the three types of elementary matrices explicitly.

a) *(Swapping two rows)* Let $i, j \in \{1, 2, \dots, m\}$ with $i \leq j$. Then

$$P_{i,j} \in M(m \times m, K), \qquad P_{i,j} := S_{i,j}(I_m)$$

is given by

$$P_{i,j} = \begin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

where the only 1's that are not on the diagonal have row-column indices (i, j) and (j, i).

b) *(Multiplying a row by λ)* Let $i \in \{1, 2, \dots, m\}$ and $\lambda \in K \setminus \{0\}$. Then

$$Q_i^\lambda \in M(m \times m, K), \qquad Q_i^\lambda := M_i^\lambda(I_m),$$

is given by

$$Q_i^\lambda = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \lambda & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1. \end{pmatrix}$$

c) *(Adding the λ-fold multiple of a row to another row)* Let $i, j \in \{1, 2, \dots, m\}$ and $\lambda \in K$. Then

$$R_{i,j}^{\lambda} \in M(m \times m, K), \qquad R_{i,j}^{\lambda} := E_{i,j}^{\lambda}(I_m)$$

is given by

$$R_{i,j}^{\lambda} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & \lambda & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

where λ is in position (i, j). (A short and more sophisticated way of writing this matrix is as

$$R_{i,j}^{\lambda} = I_m + \lambda \cdot A_{ij},$$

where $A_{ij} = (\alpha_{k\ell}) \in M(m \times m, K)$ is given by $\alpha_{k\ell} = \begin{cases} 1 & \text{if } (k, \ell) = (i, j), \\ 0 & \text{else.} \end{cases}$)

Let us now focus on quadratic matrices. Before using elementary matrices, we observe in the following result that the invertibility of a quadratic matrix can be read from its row-echelon form.

Theorem/Definition 8.61. *Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$. Then A is invertible if and only if it is row-equivalent to a matrix $B = (b_{ij}) \in M(n \times n, K)$ of the form*

$$\begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & b_{22} & b_{23} & \dots & b_{2n} \\ 0 & 0 & b_{33} & \dots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & b_{nn} \end{pmatrix}$$

i.e. with $b_{ij} = 0$ whenever $j > i$ and with $b_{ii} \neq 0$ for all $i \in \{1, 2, \dots, n\}$. A matrix which is of this form is called upper triangular.

Proof. "$\Leftarrow$": One immediately derives from its properties that $\operatorname{rank} B = n$. Hence, $\operatorname{rank} A = n$ by Theorem 8.53, which means that $\operatorname{rank} f_A = n$ and thus that $f_A : K^n \to K^n$ is surjective. Corollary 7.23.c) then implies that f_A is an isomorphism, so that A is invertible.
"$\Rightarrow$": By definition, A is invertible if and only if the map $f_A : K^n \to K^n$ is bijective which yields that $\operatorname{rank} A = \operatorname{rank} f_A = n$. By Theorem 8.57, there exists $B = (b_{ij})$ in row-echelon form and row-equivalent to A, which satisfies $\operatorname{rank} B = n$ by Theorem 8.53. Thus, for each $i \in \{1, 2, \dots, n\}$ there is a pivot element b_{ij_i} with $1 \leq j_1 < j_2 < \dots < j_n \leq n$. But the only possibility for this to happen is that $j_1 = 1$, $j_2 = 2$, $\dots$, $j_n = n$, which shows that B is of the above form. □

As a next step, we make a useful observation about elementary matrices.

Theorem 8.62. *Each elementary matrix is invertible and its inverse is again an elementary matrix. More precisely, the following holds for all $m \in \mathbb{N}$ and the elementary matrices in $M(m \times m, K)$:*

a) $P_{i,j}^{-1} = P_{i,j}$ *for all* $i, j \in \{1, 2, \dots, m\}$.

b) $(Q_i^\lambda)^{-1} = Q_i^{\frac{1}{\lambda}}$ *for all* $i \in \{1, 2, \dots, m\}$ *and* $\lambda \in K \setminus \{0\}$.

c) $(R_{i,j}^\lambda)^{-1} = R_{i,j}^{-\lambda}$ *for all* $i, j \in \{1, 2, \dots, m\}$ *and* $\lambda \in K$.

Sketch of proof. This follows from explicit computations of matrix products or from meditating about the associated elementary row operations. More generally, one might show that

$$Q_i^\lambda \cdot Q_i^\mu = Q_i^{\lambda\mu} \qquad \forall i \in \{1, 2, \dots, m\},\ \lambda, \mu \in K \setminus \{0\}$$

and

$$R_{i,j}^\lambda \cdot R_{i,j}^\mu = R_{i,j}^{\lambda+\mu} \qquad \forall i, j \in \{1, 2, \dots, m\},\ \lambda, \mu \in K.$$

We omit the details. □

The key to explicitly computing the inverse of an invertible matrix is the following observation.

Theorem 8.63. *Let $m \in \mathbb{N}$. Then every invertible $m \times m$-matrix is given as a product of finitely many elementary matrices.*

Proof. By Theorem 8.57, there exist $k \in \mathbb{N}$ and elementary matrices $E_1, \dots, E_r$, such that $B := E_1 E_2 \dots E_r A$ is an upper triangular matrix. If $B = (b_{ij})$, then we can proceed by further elementary row operations as follows:

- For each $i \in \{1, 2, \dots, m-1\}$ put
$$\lambda_{im} := -\frac{b_{im}}{b_{mm}}$$
and add the λ_{im}-fold multiple of the m-th row to the i-th row, i.e. subsequently multiply B with $R_{i,m}^{\lambda_{im}}$ and call the result
$$C_m := R_{1,m}^{\lambda_{1m}} R_{2,m}^{\lambda_{2m}} \dots R_{m-1,m}^{\lambda_{(m-1)m}} B.$$
Then the only non-zero entry in the m-th column of C_m is in position (m, m).

- Let $k \in \{2, 3, \dots, m\}$ and let $C_k = (c_{ij}^k)$ be given. For each $i \in \{1, 2, \dots, k-1\}$ put
$$\lambda_{ik} := -\frac{c_{ik}^k}{c_{kk}^k}$$
and add the λ_{ik}-fold multiple of the k-th row of C_k to the i-th row, i.e. subsequently multiply C_k with $R_{i,k}^{\lambda_{ik}}$ and call the result
$$C_{k-1} := R_{1,k}^{\lambda_{1k}} R_{2,k}^{\lambda_{2k}} \dots R_{k-1,k}^{\lambda_{(k-1)k}} C_k.$$
Then for each $j \geq k$ the only non-zero-entry in the j-th column of C_j is in position (j, j).

Following this iterative construction, one obtains a matrix C_1, which is of the form

$$C_1 = \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix},$$

where $d_i \neq 0$ for all $i \in \{1, 2, \dots, n\}$. Then

$$M_1^{d_1^{-1}} M_2^{d_2^{-1}} \dots M_m^{d_m^{-1}} C_1 = I_m.$$

Thus, by applying a (possibly large, but) finite number of elementary row operations, one can transform A into I_m, i.e. there are $k \in \mathbb{N}$ and elementary matrices $B_1, \dots, B_k$, such that

$$I_m = B_1 B_2 \dots B_k A$$

and thus by Theorem 8.15.b),

$$A = (B_1 B_2 \dots B_k)^{-1} = B_k^{-1} \dots B_2^{-1} B_1^{-1}.$$

By Theorem 8.62, each B_i^{-1} is again an elementary matrix, which shows the claim. □

Using the previous theorem and observations from its proof, we obtain a method for explicitly inverting a quadratic matrix.

Method of computing the inverse of a given invertible quadratic matrix:
Let $m \in \mathbb{N}$ and $A \in \mathrm{GL}(m, K)$.

- Apply a sequence of elementary row operations to A after which one obtains I_m.
- Apply *the same* elementary row operations to I_m and call the result B.

Since B is then given as the product of those elementary matrices which "turn A into I_m", it follows that $B = A^{-1}$.

Example 8.64. Let $A = \begin{pmatrix} 0 & 1 & -4 \\ 1 & 2 & -1 \\ 1 & 1 & 2 \end{pmatrix} \in M(3 \times 3, \mathbb{R})$. One computes that its rows are linearly independent, such that A is invertible. We compute the steps of the Gaussian algorithm for A and I_3 simultaneously:

$$\begin{pmatrix} 0 & 1 & -4 \\ 1 & 2 & -1 \\ 1 & 1 & 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\downarrow S_{1,2} \qquad \downarrow S_{1,2}$$

$$\begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & -4 \\ 1 & 1 & 2 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\begin{array}{c} \\ \downarrow E_{3,1}^{-1} \\ \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & -4 \\ 0 & -1 & 3 \end{pmatrix} \\ \downarrow E_{3,2}^{1} \\ \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & -4 \\ 0 & 0 & -1 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ \downarrow E_{3,1}^{-1} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix} \\ \downarrow E_{3,2}^{1} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} \end{array}$$

Now that the left-hand matrix is upper triangular, we apply the method from the proof of Theorem 8.63.

$$\begin{array}{c} \\ \downarrow M_3^{-1} \\ \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{pmatrix} \\ \downarrow E_{2,3}^{4} \\ \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \downarrow E_{1,3}^{1} \\ \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \downarrow E_{1,2}^{-2} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ \downarrow M_3^{-1} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ -1 & 1 & -1 \end{pmatrix} \\ \downarrow E_{2,3}^{4} \\ \begin{pmatrix} 0 & 1 & 0 \\ -3 & 4 & -4 \\ -1 & 1 & -1 \end{pmatrix} \\ \downarrow E_{1,3}^{1} \\ \begin{pmatrix} -1 & 2 & -1 \\ -3 & 4 & -4 \\ -1 & 1 & -1 \end{pmatrix} \\ \downarrow E_{1,2}^{-2} \\ \begin{pmatrix} 5 & -6 & 7 \\ -3 & 4 & -4 \\ -1 & 1 & -1 \end{pmatrix} \end{array}$$

Since we have arrived at I_3 on the left-hand side, it follows that $A^{-1} = \begin{pmatrix} 5 & -6 & 7 \\ -3 & 4 & -4 \\ -1 & 1 & -1 \end{pmatrix}$.

8.4.3 Solution sets of systems of linear equations

Finally, we are going to study solution sets of linear systems and how to determine them making use of the methods that we studied in this section. As defined in Section 7.3 a linear system or system of linear equations is a family of equations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = b_2 \\ \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots & \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = b_m. \end{cases} \tag{8.3}$$

where $a_{ij}, b_i \in K$ for all $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$. The brace indicates that we are searching for those $(x_1, \ldots, x_n) \in K^n$ which satisfy all of these equations simultaneously. Now

that we have etablished the terminology of matrices, we can rewrite such a system in a more concise way.

Definition 8.65. Consider a linear system of the form (8.3). The matrix $A := (a_{ij}) \in M(m \times n, K)$ is called *the coefficient matrix* of the linear system.

With $b := (b_1, \ldots, b_m) \in K^m$ and A being the coefficient matrix, the linear system (8.3) is equivalent to

$$Ax = b \quad \Leftrightarrow \quad f_A(x) = b,$$

where $x \in K^n$. We denote its solution set by $\mathcal{L}_{A,b} := \mathcal{L}_{f_A,b} = \{x \in K^n \mid Ax = b\}$. If $x_0 \in K^n$ satisfies $Ax_0 = b$, then by Theorem 7.30

$$\mathcal{L}_{A,b} = x_0 + \mathcal{L}_{A,0}.$$

If $T \in \mathrm{GL}(m, K)$, then obviously $Ax = b$ is equivalent to $TAx = Tb$, so

$$\mathcal{L}_{A,b} = \mathcal{L}_{TA,Tb} \qquad \forall T \in \mathrm{GL}(m, K).$$

We will make use of this observation by applying the Gaussian elimination method. A consequence of the previous discussions, in particular of Theorem 8.57 and its proof, it the following statement.

Corollary 8.66. *Let $m, n \in \mathbb{N}$. For each $A \in M(m \times n, K)$ there exists $S \in \mathrm{GL}(m, K)$, such that $S^{-1}A$ is in row-echelon form.*

This corollary can be employed to study the existence of solutions of linear equations of the form $Ax = b$. Obviously, any homogeneous equation $Ax = 0$ possesses the solution $x_0 = (0, 0, \ldots, 0)$, but what about inhomogeneous linear equations?

Theorem 8.67. *Let $A \in M(m \times n, K)$ and $b \in K^m$ and let $r = \operatorname{rank} A$. Let $S \in \mathrm{GL}(m, K)$, such that $S^{-1}A$ is in row-echelon form and put $b' := S^{-1}b$. The following are equivalent:*

(i) $Ax = b$ has a solution.

(ii) $b'_i = 0$ for all $i > r$, where $b' = (b'_1, \ldots, b'_m)$.

Proof. (i) $\Rightarrow$ (ii): The matrix $B = (b_{ij}) := S^{-1}A$ is in row-echelon form, hence $b_{ij} = 0$ for all $i > r$ and all $j \in \{1, 2, \ldots, n\}$. Since $Ax = b$ has a solution, $Bx = b'$ has a solution $x = (x_1, \ldots, x_n)$ as well. Then for $i > r$ the equality of the i-th row of $Bx = b'$ reads as

$$b'_i = \sum_{j=1}^{n} b_{ij} x_j = 0.$$

(ii) $\Rightarrow$ (i): Let $J := \{j_1, \ldots, j_r\}$ be the set of column indices of the pivots of $B := S^{-1}A$, labelled such that $j_1 < j_2 < \cdots < j_r$. Let

$$U_J := \{(x_1, \ldots, x_n) \in K^n \mid x_j = 0 \text{ if } j \notin J\} = \operatorname{span}(\{e_{j_1}, e_{j_2}, \ldots, e_{j_r}\}),$$

where the e_j denote the elements of the canonical basis of K^n. Hence U_J is a linear subspace of K^n. We want to show that the equation $Bx = b'$ has a unique solution in U_J.

If $x = \sum_{k=1}^{r} x_k e_{j_k} \in U_J$, then, since by assumption the last $m - r$ equations of the corresponding linear system are trivially satisfied, $Bx = b'$ is equivalent to

$$\begin{aligned} b_{1j_1}x_q + b_{1j_2}x_2 + \cdots + b_{1j_r}x_r &= b'_1 \\ b_{2j_2}x_2 + \cdots + b_{2j_r}x_r &= b'_2 \\ \vdots \qquad & \quad \vdots \\ b_{rj_r}x_r &= b'_r. \end{aligned}$$

Since the pivots satisfy $b_{kj_k} \neq 0$ for each $k \in \{1, 2, \ldots, r\}$, one derives that this linear system has a unique solution in the following way: The last of the above equations yields

$$x_r = \frac{b'_r}{b_{rj_r}}.$$

Inserting this into the second-to-last equation, which reads as

$$b_{(r-1)j_{r-1}}x_{r-1} + b_{(r-1)j_r}x_r = b'_{r-1}$$

we obtain a formula for x_{r-1}, which is thus uniquely determined as well. Proceeding iteratively, we compute all of the x_j and finally obtain a unique $x \in U_J$ which is a solution of $Bx = b'$ and thus of $Ax = b$. □

We not only want to confirm the existence of solutions of linear systems, but also to find *all* of their solutions. Assuming the existence of a solution, we establish the following general method for determining solution sets. As it was the case for earlier results, the proof of this theorem will provide us with a method of explicitly computing solution sets.

Theorem 8.68. *Let $A \in M(m \times n, K)$ and let $r = \operatorname{rank} A$. Let $b \in K^m$, such that $Ax = b$ admits a solution. Let B be a matrix in row-echelon form that is row-equivalent to A and let $J \subset \{1, 2, \ldots, n\}$ be the set of column indices of the pivots of B. Let $U_J, V_J \subset K^n$ be given by $U_J = \operatorname{span}(\{e_j \mid j \in J\})$ and $V_J = \operatorname{span}(\{e_j \mid j \in \{1, 2, \ldots, n\} \setminus J\})$. Then there exists $c \in K^n$ and a map $g : V_J \to c + U_J$, such that*

$$\mathcal{L}_{A,b} = \{y + g(y) \in K^n \mid y \in V_J\}.$$

Proof. Let $S \in \mathrm{GL}(m, K)$ such that $B = S^{-1}A$ and put $b' := S^{-1}b$. Then for every $x \in K^n$

$$Ax = b \quad \Leftrightarrow \quad Bx = b'.$$

Let $K^r_m := K^r \times \{0\} = \{(x_1, \ldots, x_r, 0, \ldots, 0) \in K^m \mid (x_1, \ldots, x_r) \in K^r\}$, which is obviously a linear subspace of K^m. As argued in the proof of Theorem 8.67 for each $q \in K^r_m$ there exists a unique $x \in U_J$ with $B(x) = q$, hence the restriction

$$B_J := f_B|_{U_J} : U_J \to K^r_m$$

is an isomorphism.
Let $y \in V_J$. By the row-echelon form of B, it holds that $-By \in K^r_m$. Since B_J is an isomorphism, there exists a unique $x_y \in U_J$, such that

$$x_y = -B_J^{-1}By.$$

Then $B(x_y + y) = B(x_y) + By = 0$. Thus, the following map is well-defined and linear:

$$G : V_j \to \mathcal{L}_{B,0}, \qquad G(y) = y + x_y = y - B_J^{-1}By.$$

Assume that $G(y) = 0$. Then $y = B_J^{-1}By \in U_J$, hence $y \in U_J \cap V_J = \{0\}$, which yields $y = 0$. By Theorem 7.10.b), this shows that G is injective. By the dimension formula for linear maps, $\dim \mathcal{L}_{B,0} = \dim \ker f_B = n - \operatorname{rank} f_B = n - r$. Since $|J| = r$, it holds that $\dim V_J = n - \dim U_J = n - r$ as well. Thus, by Theorem 7.10.c), G is an isomorphism. Thus,

$$\mathcal{L}_{A,0} = \mathcal{L}_{B,0} = \{y - B_J^{-1}By \in K^n \mid y \in V_J\}.$$

As discussed above, $\mathcal{L}_{A,b} = \mathcal{L}_{B,b'}$. By construction, $B_J^{-1}(b') \in \mathcal{L}_{A,b}$. Combining this with the set of homogeneous solutions,

$$\mathcal{L}_{A,b} = \{y - B_J^{-1}By + B_J^{-1}(b') \in K^n \mid y \in V_J\} = \{y + B_J^{-1}(b' - By) \in K^n \mid y \in V_J\}.$$

The claim thus follows with $c := B_J^{-1}(b')$ and $g(y) = c - B_J^{-1}By$. □

This result and the methods of its proof now lead us to the desired method of determining solutions of general systems of linear equations.

Method for finding the solution sets of systems of linear equations:

- Interpret (8.3) as an equality between matrices $Ax = b$, or explicitly

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

- Apply elementary row operations to *both A and b* to transform A into a matrix in row-echelon form B, such that $Ax = b$ is equivalent to $Bx = b'$.
- Use Theorem 8.67 to check if the system has a solution. If not, the solution set is empty.
- Find U_J and V_J as defined in Theorem 8.68.
- For each $y \in V_J$ find the unique solution $x_y \in U_J$ of $B(x_y) = -By + b'$ by Gaussian elimination and the methods of Section 8.4.2.

Then $\mathcal{L}_{A,b} = \{y + x_y \in K^n \mid y \in V_J\}$.

Example 8.69. Consider the linear system

$$\begin{cases} x_1 + 6x_3 + x_4 & = 3 \\ x_2 - 8x_3 & = -5 \\ 4x_2 - 16x_3 - 3x_4 & = -7 \end{cases}$$

which is equivalent to

$$\begin{pmatrix} 1 & 0 & 6 & 1 \\ 0 & 1 & -8 & 0 \\ 0 & 4 & -16 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3 \\ -5 \\ -7 \end{pmatrix}$$
$$E_{3,2}^{-4} \downarrow \qquad\qquad E_{3,2}^{-4} \downarrow$$
$$\begin{pmatrix} 1 & 0 & 6 & 1 \\ 0 & 1 & -8 & 0 \\ 0 & 0 & 16 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3 \\ -5 \\ 13 \end{pmatrix}$$

We call the the left-hand matrix B. Then B is in row-echelon form with $J = \{1, 2, 3\}$, hence

$$U_J = \{(x_1, x_2, x_3, 0) \in \mathbb{R}^4 \mid (x_1, x_2, x_3) \in \mathbb{R}^3\} \qquad \text{and} \qquad V_J = \{(0, 0, 0, x_4) \in \mathbb{R}^4 \mid x_4 \in \mathbb{R}\}.$$

For $y \in \mathbb{R}$ we now need to solve the equation of matrices

$$B \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix} = -B \begin{pmatrix} 0 \\ 0 \\ 0 \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ -5 \\ 13 \end{pmatrix}$$

As a linear system, this becomes

$$\begin{aligned} x_1 + 6x_3 &= -y + 3 \\ x_2 - 8x_3 &= -5 \\ 16x_3 &= 3y + 13. \end{aligned}$$

The last row yields $x_3 = \frac{3}{16}y + \frac{13}{16}$. Inserting this into the second row shows

$$x_2 = -5 + \frac{3}{2}y + \frac{13}{2} = \frac{3}{2}y + \frac{3}{2}.$$

Inserting the third into the second row yields

$$x_1 = -y + 3 - \frac{18}{16}y - \frac{78}{16} = -\frac{17}{8}y - \frac{15}{8}.$$

Consequently, the solution set of $Ax = b$ is given by

$$\begin{aligned} \mathcal{L}_{A,b} &= \Big\{ \Big(-\frac{17}{18}y - \frac{15}{18}, \frac{3}{2}y + \frac{3}{2}, \frac{3}{16}y + \frac{13}{16}, y \Big) \in \mathbb{R}^4 \;\Big|\; y \in \mathbb{R} \Big\} \\ &= \Big\{ \Big(-\frac{15}{18}, \frac{3}{2}, \frac{13}{16}, 0 \Big) + y \Big(-\frac{17}{18}, \frac{3}{2}, \frac{3}{16}, 1 \Big) \in \mathbb{R}^4 \;\Big|\; y \in \mathbb{R} \Big\}. \end{aligned}$$

Chapter 9

Properties of endomorphisms

Throughout this chapter, let K be a field.

Before discussing new mathematics, let us explain the headline of this chapter.

Definition 9.1. Let V be a vector space over K. A linear map $f : V \to V$ is called an *endomorphism of V*. A bijective linear map $f : V \to V$ is called an *automorphism*.

In other word, an endomorphism is a homomorphism for which domain and codomain are the same, while an automorphism is an endomorphism which is also an isomorphism.

9.1 Definition and construction of the determinant

In this section, we want to discuss a map which we first define by demanding it to have certain properties. We will give an explicit definition of that map below.

Definition 9.2. Let $n \in \mathbb{N}$. The *determinant* is a map

$$\det : (K^n)^n = K^n \times K^n \times \cdots \times K^n \to K$$

with the following properties:

(1) det *is multilinear:* For all $i \in \{1, 2, \ldots, n\}$, $v_1, \ldots, v_n, w_i \in K^n$ and all $\lambda, \mu \in K$ it holds that

$$\begin{aligned}&\det(v_1, \ldots, v_{i-1}, \lambda v_i + \mu w_i, v_{i+1}, \ldots, v_n)\\ &= \lambda \det v_1, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_n) + \mu \det(v_1, \ldots, v_{i-1}, w_i, v_{i+1}, \ldots, v_n).\end{aligned}$$

(2) det *is alternating:* Let $v_1, v_2, \ldots, v_n \in K^n$. If there are $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$, but $v_i = v_j$, then $\det(v_1, v_2, \ldots, v_n) = 0$.

(3) det *is normed:* $\det(e_1, e_2, \ldots, e_n) = 1$, where $(e_1, \ldots, e_n)$ is the canonical basis of K^n.

So far, the properties in Definition 9.2 are merely a wishlist. We don't know at this point if there actually exists a map which has all of these properties. Indeed, we will show in this section that there exists a *unique* map which has all of these properties and derive an explicit formula. We first make an observation that follows straight from the defining properties.

Theorem 9.3. *Let $n \in \mathbb{N}$ and $i, j \in \{1, 2, \dots, n\}$ with $i \leq j$. Then*

$$\det(v_1, \dots, v_n) = -\det(v_1, \dots, v_{i-1}, v_j, v_{i+1}, \dots, v_{j-1}, v_i, v_{j+1}, \dots, v_n) \qquad \forall v_1, \dots, v_n \in K^n.$$

Proof. By the defining properties of det, it holds that

$$\begin{aligned}
0 &\overset{(2)}{=} \det(v_1, \dots, v_{i-1}, v_i + v_j, v_{i+1}, \dots, v_{j-1}, v_i + v_j, v_{j+1}, \dots, v_n) \\
&\overset{(1)}{=} \det(v_1, \dots, v_{i-1}, v_i, \dots, v_{j-1}, v_i + v_j, \dots, v_n) + \det(v_1, \dots, v_{i-1}, v_j, \dots, v_{j-1}, v_i + v_j, \dots, v_n) \\
&\overset{(1)}{=} \det(v_1, \dots, v_{i-1}, v_i, \dots, v_{j-1}, v_i, \dots, v_n) + \det(v_1, \dots, v_{i-1}, v_i, \dots, v_{j-1}, v_j, \dots, v_n) \\
&\quad + \det(v_1, \dots, v_{i-1}, v_j, \dots, v_{j-1}, v_i, \dots, v_n) + \det(v_1, \dots, v_{i-1}, v_j, \dots, v_{j-1}, v_i, v_{j+1}, \dots, v_n) \\
&\overset{(2)}{=} \det(v_1, \dots, v_{i-1}, v_i, \dots, v_{j-1}, v_j, \dots, v_n) + \det(v_1, \dots, v_{i-1}, v_j, \dots, v_{j-1}, v_i, \dots, v_n).
\end{aligned}$$

Bringing one term to the other side shows the claim. □

Finding an explicit formula for det requires some combinatorial preparations.

Definition 9.4. Let $n \in \mathbb{N}$.

a) The *permutation group of degree n* is given by

$$S_n = \{\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\} \mid \sigma \text{ is bijective}\}.$$

together with the composition of maps. By Exercise 2 of Sheet 0 from Mathematics for Physicists 1, $(S_n, \circ)$ is a group.

b) Elements are of S_n called *permutations* of $\{1, 2, \dots, n\}$.

c) $\tau \in S_n$ is called a *transposition* if there are $i, j \in \{1, 2, \dots, n\}$, such that

$$\tau(i) = j, \qquad \tau(j) = i, \qquad \tau(k) = k \quad \forall k \in \{1, 2, \dots, n\} \setminus \{i, j\},$$

i.e. if τ swaps two elements and keeps the rest of them unchanged.

We will use the following simple fact from group theory without proving them.

Theorem 9.5. *a) Given $n \in \mathbb{N}$, the permutation group S_n has precisely $n!$ elements.*

b) Every $\sigma \in S_n$ is given as the composition of finitely many transpositions.

Definition 9.6. Let $n \in \mathbb{N}$ and $\sigma \in S_n$. If $\tau_1, \tau_2, \dots, \tau_k \in S_n$ are transpositions with

$$\sigma = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_k,$$

then we call

$$\operatorname{sign}(\sigma) := (-1)^k.$$

the *sign of σ*.

Remark 9.7. (1) In general, there are various ways of obtaining a permutation as a composition of transpositions, so it is not immediately clear that $\operatorname{sign}(\sigma)$ is well-defined. To see this, one needs to show that if $\sigma \in S_n$ and it $\tau_1, \dots, \tau_k, \rho_1, \dots, \rho_\ell \in S_n$ are transpositions, such that

$$\sigma = \tau_1 \circ \cdots \circ \tau_k = \rho_1 \circ \cdots \circ \rho_\ell,$$

then either both k and ℓ are even or both k and ℓ are odd. This is indeed the case and we take this assertion for granted in this lecture without giving a proof.

(2) It follows straight from the definition of the sign that for all $\sigma_1, \sigma_2 \in S_n$ it holds that

$$\operatorname{sign}(\sigma_1 \circ \sigma_2) = \operatorname{sign}(\sigma_1) \cdot \operatorname{sign}(\sigma_2).$$

Theorem 9.8 (Uniqueness and existence of the determinant)**.** *Let $n \in \mathbb{N}$. There exists a unique map*

$$\det : K^n \times \cdots \times K^n \to K$$

which has all three properties from Definition 9.2. If $a_1, \ldots, a_n \in K^n$ have entries $a_i = (a_{i1}, a_{i2}, \ldots, a_{in})$ for all $i \in \{1, 2, \ldots, n\}$, then

$$\det(a_1, \ldots, a_n) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \ldots a_{n\sigma(n)}. \tag{9.1}$$

Proof. Let $\det : (K^n)^n \to K$ have properties (1), (2) and (3) from Definition 9.2. Since $a_i = \sum_{j=1}^n a_{ij} e_j$ for each $i \in \{1, 2, \ldots, n\}$, we compute that

$$\begin{aligned}
\det(a_1, \ldots, a_n) &= \det\Big(\sum_{j_1=1}^n a_{1j_1} e_{j_1}, a_2, \ldots, a_n \Big) \\
&\overset{(1)}{=} \sum_{j_1=1}^n a_{1j_1} \det(e_{j_1}, a_2, \ldots, a_n) \\
&= \sum_{j_1=1}^n a_{1j_1} \det\Big(e_{j_1}, \sum_{j_2=1}^n a_{2j_2} e_{j_2}, \ldots, a_n \Big) \\
&\overset{(1)}{=} \sum_{j_1=1}^n \sum_{j_2=1}^n a_{1j_1} a_{2j_2} \det(e_{j_1}, e_{j_2}, a_3, \ldots, a_n) \\
&= \quad \ldots \\
&= \sum_{k=1}^n \sum_{j_k=1}^n a_{1j_1} a_{2j_2} \ldots a_{nj_n} \det(e_{j_1}, e_{j_2}, \ldots, e_{j_n}).
\end{aligned}$$

By property (2) of det, it holds in particular that $\det(e_{j_1}, e_{j_2}, \ldots, e_{j_r}) = 0$ if there are $k, \ell \in \{1, 2, \ldots, n\}$ with $k \neq \ell$ and $j_k = j_\ell$. Thus, the only non-vanishing summands in the above big sum occur if $j_k \neq j_\ell$ whenever $k \neq \ell$. Equivalently, the corresponding summand is non-vanishing if and only if the map $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$, $\sigma(k) = j_k$, is a permutation. Vice versa, with every permutation we can associate a unique non-vanishing summand of the above sum. We derive:

$$\det(a_1, \ldots, a_n) = \sum_{\sigma \in S_n} a_{1\sigma(1)} a_{2\sigma(2)} \ldots a_{n\sigma(n)} \det(e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}). \tag{9.2}$$

It remains to clarify the relation between $\det(e_{\sigma(1)}, \ldots, e_{\sigma(n)})$ and $\det(e_1, \ldots, e_n) \overset{(3)}{=} 1$. Let $\sigma \in S_n$ be fixed. Let $\tau, \sigma' \in S_n$ be such that τ is a transposition and such that $\sigma = \tau \circ \sigma'$. Then by Theorem 9.3

$$\det(e_{\sigma(1)}, \ldots, e_{\sigma(n)}) = \det(e_{\tau(\sigma'(1))}, \ldots, e_{\tau(\sigma'(n))}) = -\det(e_{\sigma'(1)}, \ldots, e_{\sigma'(n)}).$$

Hence, if σ is given as $\sigma = \tau_1 \circ \cdots \circ \tau_k$, where $\tau_1, \ldots, \tau_k \in S_n$ are transpositions, then

$$\det(e_{\sigma(1)}, \ldots, e_{\sigma(n)}) = (-1)^k \det(e_1, \ldots, e_n) = (-1)^k = \operatorname{sign}(\sigma).$$

Inserting this into (9.2) yields

$$\det(a_1, \dots, a_n) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}.$$

So we have shown that if a determinant map exists, it must necessarily be given by this expression. It remains to show that the map defined by this expression actually has properties (1), (2) and (3) of a determinant.

(1) (Multilinearity) Let $i \in \{1, 2, \dots, n\}$, $\lambda, \mu \in K$ and $b_i = (b_{i1}, \dots, b_{in}) \in K^n$. Then:

$$\begin{aligned}
&\det(a_1, \dots, a_{i-1}, \lambda a_i + \mu b_i, a_{i+1}, \dots, a_n) \\
&= \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{(i-1)\sigma(i-1)} (\lambda a_{i\sigma(i)} + \mu b_{i\sigma(i)}) a_{(i+1)\sigma(i+1)} \dots a_{n\sigma(n)}) \\
&= \lambda \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{(i-1)\sigma(i-1)} a_{i\sigma(i)} a_{(i+1)\sigma(i+1)} \dots a_{n\sigma(n)}) \\
&\quad + \mu \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{(i-1)\sigma(i-1)} b_{i\sigma(i)} a_{(i+1)\sigma(i+1)} \dots a_{n\sigma(n)}) \\
&= \lambda \det(a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_n) + \mu \det(a_1, \dots, a_{i-1}, b_i, a_{i+1}, \dots, a_n),
\end{aligned}$$

which we wanted to show.

(2) (Alternation) *(This part of the proof is not discussed in the lecture videos.)*
Let $i, j \in \{1, 2, \dots, n\}$ with $i \neq j$ and $a_i = a_j$, i.e. $a_{ik} = a_{jk}$ for all $k \in \{1, 2, \dots, n\}$. Let $\tau \in S_n$ be the transposition with $\tau(i) = j$ and $\tau(j) = i$. One can divide S_n into those permutations that are obtained by first applying τ and those which are not: there exists $A_n \subset S_n$ with $A_n \cap B_n = \varnothing$, where $B_n = \{\sigma \circ \tau \in S_n \mid \sigma \in A_n\}$, such that $S_n = A_n \cup B_n$. Then

$$\begin{aligned}
&\det(a_1, \dots, a_n) \\
&= \sum_{\sigma \in A_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)} + \sum_{\sigma' \in B_n} \operatorname{sign}(\sigma') a_{1\sigma'(1)} a_{2\sigma'(2)} \dots a_{n\sigma'(n)} \\
&= \sum_{\sigma \in A_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{n\sigma(n)} + \sum_{\sigma \in A_n} \operatorname{sign}(\sigma \circ \tau) a_{1\sigma(\tau(1))} \dots a_{i\sigma(\tau(i))} \dots a_{j\sigma(\tau(j))} \dots a_{n\sigma(\tau(n))} \\
&\overset{\text{Rem. 9.7.(2)}}{=} \sum_{\sigma \in A_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{n\sigma(n)} - \sum_{\sigma \in A_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} \dots a_{i\sigma(j)} \dots a_{j\sigma(i)} \dots a_{n\sigma(n)} = 0,
\end{aligned}$$

since $a_{i\sigma(i)} = a_{j\sigma(i)}$ and $a_{j\sigma(j)} = a_{i\sigma(j)}$ for all $\sigma \in S_n$ by assumption and thus $a_{1\sigma(1)} \dots a_{n\sigma(n)} = a_{1\sigma(1)} \dots a_{i\sigma(j)} \dots a_{j\sigma(i)} \dots a_{n\sigma(n)}$ for all $\sigma \in S_n$. Hence, det is alternating.

(3) (Normed map) If $a_i = e_i$ for every i, then $a_{ij} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$. Hence, for $\sigma \in S_n$ we compute

$$a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)} = \begin{cases} 1 & \text{if } \sigma(i) = i \ \forall i \in \{1, 2, \dots, n\}, \\ 0 & \text{else.} \end{cases}$$

Hence, with $\mathrm{id} := \mathrm{id}_{\{1,2,\dots,n\}} \in S_n$, the formula (9.1) becomes

$$\det(a_1, \dots, a_n) = \operatorname{sign}(\mathrm{id}) a_{11} a_{22} \dots a_{nn} = 1.$$

This shows that the map det, given by (9.1), is the unique map having all three properties of Definition 9.2. □

9.2 Determinants of matrices and endomorphisms

The determinant that we constructed in the previous section had n vectors in K^n as inputs. We can now define the determinant of an $n \times n$-matrix by simply considering its *row vectors* as inputs.

Definition 9.9. Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$. Let $(a_1^r, \dots, a_n^r)$ be the row vectors of A and view them as $a_i^r \in K^n$ for each $i \in \{1, 2, \dots, n\}$. The *determinant of* A is given by

$$\det A := \det(a_1^r, \dots, a_n^r) \in K.$$

One also writes $|A| := \det A$.

Please keep in mind that the determinant is only defined for *quadratic* matrices. Explicitly,

$$\det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}.$$

Note that each of the expressions $a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}$ contains *precisely one entry from each row and precisely one entry from each column* of the matrix.

Remark 9.10. We want to elaborate on three cases in which we can derive concise explicit formulas for $\det A$ from its description in (9.1).

(1) Let $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in M(2 \times 2, K)$. Evidently, $S_2 = \{\mathrm{id}, \tau\}$, where $\tau : \{1, 2\} \to \{1, 2\}$ is given by $\tau(1) = 2$ and $\tau(2) = 1$. Since τ is a transposition, $\operatorname{sign}(\tau) = -1$. Hence,

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

(2) Let $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \in M(3 \times 3, K)$. By Theorem 9.5.a), S_3 has precisely six elements[1]. A thorough computation of signs then shows that

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}.$$

This formula is also called the *Rule of Sarrus* and is easy to recall by a little visual memorization trick, see the explanations in the lecture video.

(3) Let $n \in \mathbb{N}$ and let $A \in M(n \times n, K)$ be *upper triangular*, so that $a_{ij} = 0$ whenever $i > j$. If $\sigma \in S_n$ is a permutation with $\sigma \neq \mathrm{id}$, then one shows that there exists an $i \in \{1, 2, \dots, n\}$ with $i > \sigma(i)$. In particular, since A is upper triangular, for each $\sigma \neq \mathrm{id}$ there exists $i \in \{1, 2, \dots, n\}$ with $a_{i\sigma(i)} = 0$. This shows that all summands but the one for id vanish in (9.1) and thus

$$\det A = a_{11}a_{22} \dots a_{nn}.$$

[1] A little fun exercise: list them all and find their signs.

To study properties of determinants of matrices, we want to investigate what effects applying elementary row operations has on the determinant of a matrix.

Theorem 9.11. *Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$.*

a) Let $i, j \in \{1, 2, \ldots, n\}$ with $i \leq j$. Then $\det(S_{i,j}(A)) = -\det A$.

b) Let $i \in \{1, 2, \ldots, n\}$ and $\lambda \in K \setminus \{0\}$. Then $\det(M_i^\lambda(A)) = \lambda \det A$.

c) Let $i, j \in \{1, 2, \ldots, n\}$ and $\lambda \in K$. Then $\det(E_{i,j}^\lambda(A)) = \det A$.

Proof. Let $(a_1^r, \ldots, a_n^r)$ be the row vectors of A.

a) By Theorem 9.3,

$$\begin{aligned}\det(S_{i,j}(A)) &= \det(a_1^r, \ldots, a_{i-1}^r, a_j^r, a_{i+1}^r, \ldots, a_{j-1}^r, a_i^r, a_{j+1}^r, \ldots, a_n^r)\\ &= -\det(a_1^r, \ldots, a_n^r) = -\det A.\end{aligned}$$

b) By the multilinearity of det,

$$\begin{aligned}\det(M_i^\lambda(A)) &= \det(a_1^r, \ldots, a_{i-1}^r, \lambda a_i^r, a_{i+1}^r, \ldots, a_n^r)\\ &= \lambda \det(a_1^r, \ldots, a_{i-1}^r, a_i^r, a_{i+1}^r, \ldots, a_n^r) = \lambda \det A.\end{aligned}$$

c) Since det is multilinear and alternating,

$$\begin{aligned}\det(E_{i,j}^\lambda(A)) &= \det(a_1^r, \ldots, a_{i-1}^r, a_i^r + \lambda a_j^r, a_{i+1}^r, \ldots, a_n^r)\\ &= \det(a_1^r, \ldots, a_{i-1}^r, a_i^r, a_{i+1}^r, \ldots, a_n) + \lambda \det(a_1^r, \ldots, a_{i-1}^r, a_j^r, a_{i+1}^r, \ldots, a_j^r, \ldots, a_n^r)\\ &= \det A + \lambda \cdot 0 = \det A.\end{aligned}$$

□

So far, we have talked about invertible matrices and observed that not every quadratic matrix is invertible. We can even invert a matrix if we know that it is invertible, but we have not yet found a method of finding out whether a given matrix is invertible or not. The following theorem shows that we can do so by using the determinant.

Theorem 9.12. *Let $n \in \mathbb{N}$ and let $A \in M(n \times n, K)$. The following statements are equivalent:*

(i) A is invertible.

(ii) $\operatorname{rank} A = n$.

(iii) $\det A \neq 0$.

Proof. The equivalence of (i) and (ii) was already discussed in the previous chapters, so it suffices to show the implications "(i) $\Rightarrow$ (iii)" and "(iii) $\Rightarrow$ (ii)".
(i) $\Rightarrow$ (iii): By Theorem/Definition 8.61, we can transform A by elementary row operations into an upper triangular matrix $B = (b_{ij})$with $b_{ii} \neq 0$ for all $i \in \{1, 2, \ldots, n\}$. By Remark 9.10,

$$\det B = b_{11} b_{22} \ldots b_{nn} \neq 0.$$

Since B is obtained from A by elementary row operations, Theorem 9.11 yields $\det A = \mu \det B$ for some $\mu \in K \setminus \{0\}$, hence $\det A \neq 0$ as well.

(iii) $\Rightarrow$ (ii): We will prove the claim by contraposition. Assume that $\operatorname{rank} A < n$. By Theorem 8.53, there is a matrix $B = (b_{ij})$ in row-echelon form obtained from A via elementary row operations with $\operatorname{rank} B = \operatorname{rank} A < n$. As in the previous implication, $\det A = \mu \det B$ for some $\mu \in K$. But since B is in row-echelon form and $\operatorname{rank} B < n$, its n-th row vanishes entirely, i.e. $b_{nj} = 0$ for all $j \in \mathbb{N}$. But since every summand of $\det B$ in the explicit formula (9.1) contains a factor of the form b_{nj}, this implies that $\det B = 0$. Hence, $\det A = 0$. By contraposition, this shows that $\det A \neq 0$ yields $\operatorname{rank} A = n$. □

We want to study determinants of matrix products. Before carrying on, we habe to insert an important result about the matrix product that the lecturer simply *forgot* to include into Section 8.1 where it actually belongs...

Theorem 9.13. *Let $\ell, m, n \in \mathbb{N}$. Let $A, B \in M(\ell \times m, K)$ and $C, D \in M(m \times n, K)$. Then*

$$A(C + D) = AC + AD \qquad \text{and} \qquad (A + B)C = AC + BC.$$

Proof. We will only prove the first equation, since the second one is proven analogously. Let $f_A, f_B : K^m \to K^\ell$ and $f_C, f_D : K^n \to K^m$ be the linear maps induces by A, B, C and D, respectively. It follows from the definition, that $C + D$ induces the map $f_C + f_D$, so by Theorem 8.9, the matrix $A(C + D)$ induces the map $f_A \circ (f_C + f_D)$. For each $x \in K^n$ we compute by linearity of f_A:

$$(f_A \circ (f_C + f_D))(x) = f_A(f_C(x) + f_D(x)) = f_A(f_C(x)) + f_A(f_D(x)) = (f_A \circ f_C)(x) + (f_A \circ f_D)(x).$$

Hence, $f_A \circ (f_C + f_D) = f_A \circ f_C + f_A \circ f_D$ and since the right-hand map is induced by $AC + AD$, this shows the claim. □

It turns out that the determinant behaves well with respect to products of matrices, which will turn out to be immensely useful.

Theorem 9.14. *For all $A, B \in M(n \times n, K)$ it holds that*

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

Proof. (We will *not* prove this theorem by an explicit computation, but use the uniqueness property of the determinant in a smart way.)
Let $B \in M(n \times n, K)$ be fixed. If B is not invertible, then a standard computation shows that $A \cdot B$ is not invertible for any $A \in M(n \times n, K)$, so in that case both side of the equation vanish. If B is invertible, then $\det(B) \neq 0$ and we define

$$d_B : M(n \times n, K) \to K, \qquad d_B(A) := \frac{\det(AB)}{\det B}.$$

We want to show that d_B has all three defining properties of the determinant. Let $A \in M(n \times n, K)$ and let $(a_1, \ldots, a_n)$ be its row vectors. A standard computation of matrix products then shows that the row vectors of AB are given by

$$(a_1 \cdot B, \ldots, a_n \cdot B).$$

(1) Let $\lambda, \mu \in K$ and let $b_i \in M(1 \times n, K)$. Then

$$\begin{aligned} &d_B(a_1, \ldots, a_{i-1}, \lambda a_i + \mu b_i, a_{i+1}, \ldots, a_n) \\ &= \frac{1}{\det B} \det(a_1 \cdot B, \ldots, a_{i-1} \cdot B, (\lambda a_i + \mu b_i) \cdot B, a_{i+1} \cdot B, \ldots, a_n \cdot B). \end{aligned}$$

By property (1) of det, this yields

$$\begin{aligned} &d_B(a_1, \ldots, a_{i-1}, \lambda a_i + \mu b_i, a_{i+1}, \ldots, a_n) \\ &= \frac{\lambda}{\det B} \det(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n) + \frac{\mu}{\det B} \det(a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n) \\ &= \lambda d_B(a_1, \ldots, a_n) + \mu d_B(a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n). \end{aligned}$$

This shows that d_B has the multilinearity property of the determinant.

(2) Let $A \in M(n \times n, K)$ and let $(a_1, \ldots, a_n)$ be its row vectors and assume that there are $i \neq j$ with $a_i = a_j$. Then $a_i \cdot B = a_j \cdot B$, so it follows from property (2) of det that

$$d_B(a_1, \ldots, a_n) = \frac{\det(a_1 \cdot B, \ldots, a_n \cdot B)}{\det B} = 0.$$

(3) We compute that $d_B(e_1, \ldots, e_n) = \frac{\det(I_n \cdot B)}{\det B} = \frac{\det B}{\det B} = 1$, so d_B is normed.

Since d_B has all three properties from Definition 9.2, it follows from the uniqueness part of Theorem 9.8 that $d_B(A) = \det A$ and thus $\det(AB) = \det(AB)$ for all $A \in M(n \times n, K)$. □

Corollary 9.15. *a) Let $A \in \mathrm{GL}(n, K)$. Then $\det(A^{-1}) = (\det A)^{-1}$. F*

b) Let $A, B \in M(n \times n, K)$ and assume there exists $P \in \mathrm{GL}(n, K)$, such that

$$B = PAP^{-1}.$$

Then $\det A = \det B$.

Proof. a) Let again I_n be the identity matrix of size n. Since I_n is upper triangular, we derive from Remark 9.10.(3) that $\det I_n = 1$. Since A is invertible, $\det A \neq 0$ by Theorem 9.12. Since $AA^{-1} = I_n$, we derive from Theorem 9.14 that

$$1 = \det I_n = \det(A \cdot A^{-1}) = \det A \cdot \det(A^{-1}) \qquad \Leftrightarrow \det A^{-1} = \frac{1}{\det A}.$$

b) Applying Theorem 9.14 twice yields

$$\det B = \det(P \cdot A \cdot P) = \det P \cdot \det A \cdot \det(P^{-1}) \overset{\text{a)}}{=} \det P \cdot \det A \cdot (\det P)^{-1} = \det A.$$

□

Corollary/Definition 9.16. *Let $\mathrm{SL}(n, K) := \{A \in M(n \times n, K) \mid \det A = 1\}$. Then $\mathrm{SL}(n, K)$ is a group with respect to multiplication of matrices. It is called* the special linear group of degree n over K.

Proof. By Theorem 9.12, $\mathrm{SL}(n, K) \subset \mathrm{GL}(n, K)$. By Theorem 9.14, it holds for all $A, B \in \mathrm{SL}(n, K)$ that $\det(AB) = (\det A)(\det B) = 1$, hence $AB \in \mathrm{SL}(n, K)$. By Corollary 9.15.a) it further holds that $\det(A^{-1}) = (\det A)^{-1} = 1$ for all $A \in \mathrm{SL}(n, K)$. Moreover, $\det I_n = 1$, hence $I_n \in \mathrm{SL}(n, K)$. The claim then follows in analogy with Corollary 8.16. □

Theorem 9.17. *Let $n \in \mathbb{N}$. For all $A \in M(n \times n, K)$ it holds that*

$$\det A = \det(A^t).$$

Proof. (This proof is not discussed in the lecture videos.)

Let $A = (a_{ij})$. By definition of A^t, we compute that

$$\det(A^t) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \dots a_{\sigma(n)n}.$$

For each $\sigma \in S_n$ it holds that

$$a_{\sigma(1)1} a_{\sigma(2)2} \dots a_{\sigma(n)n} = a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} \dots a_{n\sigma^{-1}(n)}, \tag{9.3}$$

since if $\sigma(k) = i_k$ for each $k \in \{1, 2, \dots, n\}$, then $\sigma^{-1}(i_k) = k$, such that $a_{\sigma(k)k} = a_{i_k k} = a_{i_k \sigma^{-1}(i_k)}$. Since every $a_{i_k \sigma^{-1}(i_k)}$ appears exactly once in the right-hand side of (9.3), this shows the validity of that equation. Putting $\tau := \sigma^{-1}$ in each summand, we obtain

$$\begin{aligned} \det(A^t) &= \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} \dots a_{n\sigma^{-1}(n)} \\ &= \sum_{\tau \in S_n} \operatorname{sign}(\tau^{-1}) a_{1\tau(1)} a_{2\tau(2)} \dots a_{n\tau(n)}. \end{aligned}$$

Since $\tau \circ \tau^{-1} = \mathrm{id}$ and since $\operatorname{sign}(\mathrm{id}) = 1$, we obtain from Remark 9.7.(2) that $\operatorname{sign}(\tau) = \operatorname{sign}(\tau^{-1})$ for each $\tau \in S_n$. Inserting this into the above shows that $\det(A^t) = \det A$. □

As a next aim, we want to define the determinant of an endomorphism of a finite-dimensional vector space. Such a determinant will be defined as the determinant of one of its matrix representations. To make this well-defined, we will first establish using the previous observations and our knowledge on matrix representations that this determinant is independent of the choice of a basis.

Definition 9.18. Let $n \in \mathbb{N}$ and $A, B \in M(n \times n, K)$. A and B are called *similar* (or *conjugate*) if there exists $P \in \mathrm{GL}(n, K)$ with $B = PAP^{-1}$.

Let $n \in \mathbb{N}$, let V be an n-dimensional vector space over K and let $f : V \to V$ be an endomorphism. Given an ordered basis B of V we can study the matrix representation

$$M_B(f) := M_{B,B}(f) \in M(n \times n, K).$$

of f, i.e. we can choose the coordinate representation with respect to the same basis on domain and target of f.

Theorem 9.19. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. If B and C are ordered bases of V, then $M_B(f)$ and $M_C(f)$ are similar.*

Proof. Let $n := \dim V$ and let $T_{B,C} \in \mathrm{GL}(n, K)$ be the transformation matrix from B to C. By Theorem 8.26, it then holds that

$$M_C(f) = M_{C,C}(f) = T_{B,C} \cdot M_{B,B}(f) \cdot T_{B,C}^{-1} = P(M_B(f))P^{-1}$$

with $P = T_{B,C}$. Hence, $M_B(f)$ and $M_C(f)$ are similar. □

Definition 9.20. Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. Given an ordered basis B of V, we put

$$\det f := \det(M_B(f))$$

and call it *the determinant of f*. (By Corollary 9.15.b) and Theorem 9.19, this number is independent of the choice of B, hence well-defined.)

The properties of determinants of matrices that we discussed in this section evidently have consequences for endomorphisms. The following theorem summarizes all of these consequences. Its proof traces them back to the corresponding statements for matrices.

Theorem 9.21. *Let V be a finite-dimensional vector space over K and let $f, g \in L(V, V)$. Then:*

a) $\det(\mathrm{id}_V) = 1$.

b) f is an automorphism if and only if $\det f \neq 0$.

c) $\det(g \circ f) = (\det g)(\det f)$.

d) Let $f^* : V^* \to V^*$ be the transpose of f. Then $\det(f^*) = \det f$.

Proof. (This proof is not carried out in the lecture videos.)

a) Let $n = \dim V$. For every ordered basis B of V, it holds that

$$M_B(\mathrm{id}_V) = M_{B,B}(\mathrm{id}_V) = I_n.$$

Hence, $\det(\mathrm{id}_V) = \det(I_n) = 1$.

b) f is an automorphism if and only if it is surjective, i.e. if $\operatorname{rank} f = n$. It follows from Corollary 8.31 that f is an automorphism if and only if $\operatorname{rank} M_B(f) = n$ for every ordered basis B of V. The claim thus follows from Theorem 9.12.

c) Let B be an ordered basis of V. By Theorem 8.27, it holds that

$$M_B(g \circ f) = M_B(g) \cdot M_B(f).$$

The claim thus follows from Theorem 9.14.

d) This follows by combining Theorem 8.48 with Theorem 9.17.

□

9.3 Computations using determinants of matrices

There are certain situations in which we can tackle some computational problems studied in Section 8.4 like solving linear systems or computing the inverse of a matrix using determinants instead of Gaussian elimination. In this section we want to study two such methods and also give an approach to computing determinants of large matrices by computing determinants of smaller ones.

Theorem 9.22 (Cramer's Rule)**.** *Let $A \in \mathrm{GL}(n, K)$ and $b \in K^n$. Let $(a_1, \ldots, a_n)$ be the* column *vectors of A. For each $i \in \{1, 2, \ldots, n\}$ let $A_i^b \in M(n \times n, K)$ be the matrix whose column vectors are*

$$(a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n),$$

i.e. the matrix we obtain from A by replacing its i-th column by b. Then the unique solution $x = (x_1, \ldots, x_n)$ of $Ax = b$ is given by

$$x_i = \frac{\det A_i^b}{\det A} \qquad \forall i \in \{1, 2, \ldots, n\}.$$

Proof. Since A is invertible, $Ax = b$ has a unique solution given by $x = A^{-1}b$. Viewing b as a column vector, we obtain

$$\sum_{j=1}^{n} x_j a_j = Ax = b.$$

Using the multilinearity property of det and Theorem 9.17, we compute for each $i \in \{1, 2, \ldots, n\}$ that

$$\begin{aligned}\det A_i^b &= \det(A_i^b)^t = \det(a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n) \\ &= \det\Big(a_1, \ldots, a_{i-1}, \sum_{j=1}^{n} x_j a_j, a_{i+1}, \ldots, a_n\Big) \\ &= \sum_{j=1}^{n} x_j \det(a_1, \ldots, a_{i-1}, a_j, a_{i+1}, \ldots, a_n).\end{aligned}$$

By the alternating property (2) of det, all but one of the summands in the above sum vanish and we obtain:

$$\det A_i^b = x_i \det(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n) = x_i \det A^t = x_i \det A.$$

Since A is invertible, $\det A \neq 0$ by Theorem 9.12 and the claim immediately follows. □

Remark 9.23. For large $n \in \mathbb{N}$, it turns out that applying Cramer's rule to solve a system of linear equations will take more computational steps than applying Gaussian elimination as in Section 8.4.3.

In the same spirit, we can use determinants to derive an alternative approach to computing the inverse of an invertible matrix.

Theorem 9.24. *Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$. Let $(a_1, \ldots, a_n)$ be the* row *vectors of A. Then*

$$A \cdot \widetilde{A} = (\det A) I_n,$$

where $\widetilde{A} = (\tilde{a}_{ij})$ is given by

$$\tilde{a}_{ij} = \det(a_1, \ldots, a_{j-1}, e_i, a_{j+1}, \ldots, a_n) \qquad \forall i, j \in \{1, 2, \ldots, n\}.$$

In particular, if A is invertible, then

$$A^{-1} = \frac{1}{\det A} \cdot \widetilde{A}.$$

Proof. Using the alternating property of det, we compute for all $i, j \in \{1, 2, \ldots, n\}$ that

$$\det(a_1, \ldots, a_{j-1}, a_i, a_{j+1}, \ldots, a_n) = \begin{cases} \det A & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} = \det A \cdot \delta_{ij}.$$

Since $a_i = \sum_{k=1}^{n} a_{ik} e_k$, where we view the e_k as row vectors, we derive from the multilinearity of det that

$$\begin{aligned}\det A \cdot \delta_{ij} &= \det\Big(a_1, \ldots, a_{j-1}, \sum_{k=1}^{n} a_{ik} e_k, a_{j+1}, \ldots, a_n\Big) \\ &= \sum_{k=1}^{n} a_{ik} \det(a_1, \ldots, a_{j-1}, e_k, a_{j+1}, \ldots, a_n) = \sum_{k=1}^{n} a_{ik} \tilde{a}_{kj}.\end{aligned}$$

By definition of the matrix product, this last sum is the (i,j)-entry of $A \cdot \widetilde{A}$. Since $I_n = (\delta_{ij})$, this yields $(\det A) \cdot I_n = A \cdot \widetilde{A}$. If A is invertible, then $\det A \neq 0$ by Theorem 9.12 and we obtain $A \cdot (\frac{1}{\det A} \cdot \widetilde{A}) = I_n$, which show by the uniqueness of inverses that $A^{-1} = \frac{1}{\det A} \cdot \widetilde{A}$. □

In the following, we want to work out a nicer formulation of the coeficients $\tilde{a}_{ij}$ in the situation of Theorem 9.24. We first establish some additional formalism.

Definition 9.25. Let $n \in \mathbb{N}$ and $A = (a_{ij}) \in M(n \times n, K)$. For $i, j \in \{1, 2, \ldots, n\}$ we define

$$A_{i,j} \in M((n-1) \times (n-1), K), \quad A_{i,j} = \begin{pmatrix} a_{11} & \cdots & a_{1(j-1)} & a_{1(j+1)} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{(i-1)1} & \cdots & a_{(i-1)(j-1)} & a_{(i-1)(j+1)} & \cdots & a_{(i-1)n} \\ a_{(i+1)1} & \cdots & a_{(i+1)(j-1)} & a_{(i+1)(j+1)} & \cdots & a_{(i+1)n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n(j-1)} & a_{n(j+1)} & \cdots & a_{nn} \end{pmatrix},$$

i.e. $A_{i,j}$ is the matrix that we obtain from A by deleting the i-th row and the j-th column of A.

Lemma 9.26. *In the situation of Theorem 9.24, it holds for all $i, j \in \{1, 2, \ldots, n\}$ that*

$$\tilde{a}_{ij} = (-1)^{i+j} \det A_{j,i}.$$

Proof. (This proof is not discussed in the lecture videos.)

By definition and by Theorem 9.11.a),

$$\tilde{a}_{ij} = \det \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{(j-1)1} & \cdots & a_{(j-1)i} & \cdots & a_{(j-1)n} \\ 0 & \cdots & 1 & \cdots & 0 \\ a_{(j+1)1} & \cdots & a_{(j+1)i} & \cdots & a_{(j+1)n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{ni} & \cdots & a_{nn} \end{pmatrix} = (-1)^{j-1} \det \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \\ a_{11} & \cdots & a_{1i} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{(j-1)1} & \cdots & a_{(j-1)i} & \cdots & a_{(j-1)n} \\ a_{(j+1)1} & \cdots & a_{(j+1)i} & \cdots & a_{(j+1)n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{ni} & \cdots & a_{nn} \end{pmatrix}.$$

Applying Theorem 9.17 and Theorem 9.11 then shows

$$\tilde{a}_{ij} = (-1)^{j-1} \det \begin{pmatrix} 0 & a_{11} & \cdots & a_{(j-1)1} & a_{(j+1)1} & \cdots & a_{n1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 1 & a_{1i} & \cdots & a_{(j-1)i} & a_{(j+1)i} & \vdots & a_{ni} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & a_{1n} & \cdots & a_{(j-1)n} & a_{(j+1)n} & \cdots & a_{nn} \end{pmatrix}$$

$$= (-1)^{i+j-2} \det \begin{pmatrix} 1 & a_{1i} & \dots & a_{(j-1)i} & a_{(j+1)i} & \dots & a_{ni} \\ 0 & a_{11} & \dots & a_{(j-1)1} & a_{(j+1)1} & \dots & a_{n1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{1(i-1)} & \dots & a_{(j-1)(i-1)} & a_{(j+1)(i+1)} & \dots & a_{n(i-1)} \\ 0 & a_{1(i+1)} & \dots & a_{(j-1)(i+1)} & a_{(j+1)(i+1)} & \dots & a_{n(i+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{1n} & \dots & a_{(j-1)n} & a_{(j+1)n} & \dots & a_{nn} \end{pmatrix}$$
$$= (-1)^{i+j} \det \begin{pmatrix} 1 & a' \\ 0 & A^t_{j,i} \end{pmatrix}.$$

Here, the latter is a block matrix with

$$a' := \begin{pmatrix} a_{1i} & \dots & a_{(j-1)i} & a_{(j+1)i} & \dots & a_{ni} \end{pmatrix} \in M(1 \times (n-1), K).$$

Using the explicit form of the determinant in (9.1), one realizes that every summand which has an entry of a' among its factors vanishes as it must an entry in position $(k, 1)$ for some $k > 1$. From the remaining summand and Theorem 9.17, one computes that $\det \begin{pmatrix} 1 & a' \\ 0 & A^T_{j,i} \end{pmatrix} = \det A_{j,i}$, which shows the claim. □

Corollary 9.27. *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2, K)$. *Then*

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Proof. In this situation, we compute that $A_{1,1}, A_{1,2}, A_{2,1}, A_{2,2} \in M(1 \times 1, K)$ are given by

$$A_{1,1} = (d), \qquad A_{1,2} = (c), \qquad A_{2,1} = (b), \qquad A_{2,2} = (a).$$

By Theorem 9.24 and Lemma 9.26, we obtain

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} (-1)^{1+1} \det A_{1,1} & (-1)^{1+2} \det A_{2,1} \\ (-1)^{2+1} \det A_{2,1} & (-1)^{2+2} \det A_{2,2} \end{pmatrix} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

□

An important tool to compute determinants of large matrices by an iterative method from the determinants of smaller matrices is the following theorem.

Theorem 9.28 (Laplace expansion theorem)**.** *Let* $n \in \mathbb{N}$ *and* $A \in M(n \times n, K)$. *Then for all* $i \in \{1, 2, \dots, n\}$*:*

a) (Expansion by the i-th row) $\det A = \sum_{k=1}^{n} (-1)^{i+k} a_{ik} \det A_{i,k}$.

b) (Expansion by the i-th column) $\det A = \sum_{k=1}^{n} (-1)^{i+k} a_{ki} \det A_{k,i}$.

Proof. a) We obtain from Theorem 9.24 and Lemma 9.26 that for each $i \in \{1, 2, \dots, n\}$:

$$\det A = \det A \cdot \delta_{ii} = \sum_{k=1}^{n} a_{ik} \tilde{a}_{ki} = \sum_{k=1}^{n} (-1)^{i+k} a_{ik} \det A_{i,k}.$$

b) Since $\det A = \det A^T$ by Theorem 9.17, this follows from part a) applied to A^T instead of A. □

Example 9.29. We want to find the determinant of $A = \begin{pmatrix} 2 & 5 & 1 & -2 \\ 1 & 0 & 3 & -2 \\ 4 & -1 & -2 & -1 \\ 3 & -5 & 2 & 4 \end{pmatrix} \in M(4 \times 4, \mathbb{R})$ by expansion by the second row. By the Laplace expansion theorem,

$$\begin{aligned} \det A &= (-1)^{2+1} 1 \cdot \det A_{2,1} + (-1)^{2+2} 0 \cdot \det A_{2,2} + (-1)^{2+3} 3 \cdot \det A_{2,3} + (-1)^{2+4} (-2) \cdot \det A_{2,4} \\ &= -\det A_{2,1} - 3 \det A_{2,3} - 2 \det A_{2,4}. \end{aligned}$$

Here,

$$A_{2,1} = \begin{pmatrix} 5 & 1 & -2 \\ -1 & -2 & -1 \\ -5 & 2 & 4 \end{pmatrix}, \qquad A_{2,3} = \begin{pmatrix} 2 & 5 & -2 \\ 4 & -1 & -1 \\ 3 & -5 & 4 \end{pmatrix}, \qquad A_{2,4} = \begin{pmatrix} 2 & 5 & 1 \\ 4 & -1 & -2 \\ 3 & -5 & 2 \end{pmatrix}.$$

Using the Rule of Sarrus, one computes that

$$\det A_{2,1} = 3, \qquad \det A_{2,3} = -79, \qquad \det A_{2,4} = -111.$$

Hence,

$$\det A = -3 + 3 \cdot 79 + 2 \cdot 111 = -3 + 237 + 222 = 456.$$

(This might be a surprisingly huge number on first sight, but it shows that the determinant of a matrix is not at all easy to estimate from its entries.)

9.4 Eigenvalues and diagonalizability

In Theorem 8.33, we have seen that every linear map $f : V \to W$ between finite-dimensional vector spaces has a matrix representation of the form

$$M_{B,C}(f) = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix},$$

where $r = \operatorname{rank} f$, for suitably chosen ordered bases B of V and C of W.
If we study an endomorphism $f : V \to V$, we would of course prefer to use *the same* coordinate representation, i.e. *the same* basis, on the domain and the codomain of f. In other words, we would prefer matrix representations of f of the form $M_B(f) = M_{B,B}(f)$ as we already did in Section 9.2. Since it is important in the proof of Theorem 8.33 to have flexibility in our choices of B and C, it is not at all clear if there exists an ordered basis B of V, for which $M_B(f)$ has a particularly simple form. Our main objective for the rest of this chapter is to investigate this question and to find criteria on f to admit a matrix representation $M_B(f)$ of a particularly simple form. Before we make this more precise, we introduce some additional terminology for endomorphisms.

Definition 9.30. Let V be a vector space over K and let $f : V \to V$ be an endomorphism. A number $\lambda \in K$ is called[2] an *eigenvalue of* f if there exists $v \in V$ with $v \neq 0$, such that

$$f(v) = \lambda v.$$

[2]The term *eigenvalue* is probably one of the strangest in all of mathematics. It is a semi-translation of the German word *Eigenwert*, which loosely translates as "own value", a value that belongs to the endomorphism itself.

The vector v is then called an *eigenvector of f corresponding to λ*. The set

$$E_\lambda(f) = \{v \in V \mid f(v) = \lambda v\}$$

is called the *eigenspace of f associated with λ*.

Remark 9.31. Let V be a vector space over K and let $f \in L(V, V)$.

(1) Evidently, 0 is an eigenvalue of f if and only if $\ker f \neq \{0\}$. Then $E_0(f) = \ker f$.

(2) Let $\lambda \in K$ be an eigenvalue of f. Then, by Theorem 7.4, $f - \lambda \mathrm{id}_V : V \to V$ is an endomorphism as well and by definition

$$E_\lambda(f) = \ker(f - \lambda \mathrm{id}_V).$$

In particular, $E_\lambda(f)$ is a linear subspace of V.

Example 9.32. Let V be a vector space over K.

(1) Let $a \in K$ and let $f : V \to V$, $f(v) = av$, which is an endomorphism. Then a is an eigenvalue of f with $E_a(f) = V$.

(2) Let $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x) = (2x, -4y)$. Evidently $f = f_A$ is linear, where $A = \begin{pmatrix} 2 & 0 \\ 0 & -4 \end{pmatrix}$. We observe that 2 is an eigenvalue of f, since $f(1,0) = (2,0) = 2 \cdot (1,0)$, and one checks that $E_2(f) = \mathbb{R} \times \{0\}$. Also, -4 is an eigenvalue of f, since $f(0,1) = (0,-4) = -4 \cdot (0,1)$ and one checks that $E_{-4}(f) = \{0\} \times \mathbb{R}$.

(3) Let $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = (y,-x)$. Then f is an endomorphism. If $\lambda \in \mathbb{R}$ was an eigenvalue of f, then

$$f(x,y) = \lambda(x,y) \quad \Leftrightarrow \quad (y,-x) = (\lambda x, \lambda y) \quad \Leftrightarrow \quad \begin{cases} y = \lambda x, \\ -x = \lambda y. \end{cases}$$

Combining the two equations yields $x = -\lambda^2 x$ and $y = -\lambda^2 y$, which only has the trivial solution $x = y = \lambda = 0$. Hence, f does *not* have any eigenvalues.

(4) Given an interval $I \subset \mathbb{R}$ we put

$$C^\infty(I, \mathbb{R}) := \bigcap_{k \in \mathbb{N}} C^k(I, \mathbb{R}) = \{f : I \to \mathbb{R} \mid f \text{ is } k \text{ times differentiable for every } k \in \mathbb{N}\}.$$

$C^\infty(I, \mathbb{R})$ is a linear subspace of $C^0(I, \mathbb{R})$, hence a real vector space. The map $D : C^\infty(\mathbb{R}, \mathbb{R}) \to C^\infty(\mathbb{R}, \mathbb{R})$, $D(f) = f'$, is an endomorphism, see Example 7.3.(4). By Corollary 4.25, *every* $\lambda \in \mathbb{R}$ is an eigenvalue of D with

$$E_\lambda(D) = \{f_a(x) = ae^{\lambda x} \mid a \in \mathbb{R}\}.$$

Our motivation for the study of eigenvalues was that we aim for matrix representations of endomorphisms that take a particularly simple form. This will be made more precise in the following definition.

Definition 9.33. Let $n \in \mathbb{N}$.

a) $A = (a_{ij}) \in M(n \times n, K)$ is called a *diagonal matrix* if $a_{ij} = 0$ whenever $i \neq j$, i.e. if A is of the form

$$A = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}.$$

b) Let V be an n-dimensional vector space over K and let $f : V \to V$ be an endomorphism. We call f *diagonizable* if there exists an ordered basis B of V, for which $M_B(f)$ is a diagonal matrix.

Theorem/Definition 9.34. *Let V be a finite-dimensional vector space over K. An endomorphism $f : V \to V$ is diagonalizable if and only if there exists a basis of V which consists of eigenvectors of f. Such a basis is called an* eigenbasis for f.

Proof. This is evident, since one derives straight from Theorem 8.20 that for an ordered basis B of V the matrix $M_B(f)$ is diagonal if and only if B consists of eigenvectors of f. □

Example 9.35. The map from Example 9.32.(2) is obviously diagonalizable, while the map from Example 9.32.(3) is not diagonalizable by Theorem/Definition 9.34, since it does not have an eigenbasis.

Theorem 9.36. *Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$. Then $f_A : K^n \to K^n$ is diagonalizable if and only if A is similar to a diagonal matrix.*

Proof. "⇒": Let $B = \{b_1, \dots, b_n\} \subset K^n$ be an eigenbasis for f_A, which exists by Theorem/Definition 9.34. For each $i \in \{1, 2, \dots, n\}$ let $\lambda_i \in K$ be the eigenvalue with $b_i \in E_{\lambda_i}(f)$. Let $S \in \mathrm{GL}(n, K)$ be the matrix whose inverse S^{-1} has column vectors $(b_1, \dots, b_n)$. Then $S^{-1}e_i = b_i$ and thus $Sb_i = e_i$ for each $i \in \{1, 2, \dots, n\}$. Consequently,

$$SAS^{-1}e_i = SAb_i = S(\lambda_i b_i) = \lambda_i Sb_i = \lambda_i e_i \qquad \forall i \in \{1, 2, \dots, n\}.$$

Hence, $SAS^{-1} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$, which is a diagonal matrix.

"⇐": Let $S \in \mathrm{GL}(n, K)$, such that SAS^{-1} is a diagonal matrix and let $(b_1, \dots, b_n)$ be the columns of S^{-1}. Let $d_1, \dots, d_n \in K$ be the entries of the diagonal of SAS^{-1}. Then

$$f_A(b_i) = AS^{-1}e_i = S^{-1}(SAS^{-1})e_i) = S^{-1}(d_i e_i) = d_i S^{-1}e_i = d_i b_i.$$

Hence, b_i is an eigenvector corresponding to d_i for each $i \in \{1, 2, \dots, n\}$, such that $\{b_1, \dots, b_n\}$ is an eigenbasis for f_A. □

To state a criterion for diagonizability, we will first generalize the notion of a direct sum of linear subspaces from two subspaces to finitely many.

Definition 9.37. Let V be a vector space over K and let $U_1, \dots, U_n \subset V$ be linear subspaces. We call the sum $U_1 + U_2 + \dots + U_n$ the *direct sum of* $U_1, \dots, U_n$ if the following holds:
If $u_1 \in U_1$, $u_2 \in U_2$, …, $u_n \in U_n$ satisfy $u_1 + u_2 + \dots + u_n = 0$, then

$$u_1 = u_2 = \dots = u_n = 0.$$

Remark 9.38. (1) For $n = 2$ the notion of direct sum from Definition 9.37 coincides with the one introduced in Definition 6.36.

(2) Let V be a vector space over K and let $U_1, \dots, U_n \subset V$ be linear subspaces. If $U_1 \oplus \cdots \oplus U_n$ is a direct sum, then

$$\dim(U_1 \oplus U_2 \oplus \cdots \oplus U_n) = \sum_{i=1}^{n} \dim U_i.$$

This follows by induction over n using Corollary 6.41 for the induction step.

Lemma 9.39. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. If $\lambda_1, \dots, \lambda_k \in K$ are pairwise distinct eigenvalues of f, i.e. with $\lambda_i \neq \lambda_j$ whenever $i \neq j$, then the eigenspaces $E_{\lambda_1}(f), \dots, E_{\lambda_k}(f)$ form a direct sum*

$$E_{\lambda_1}(f) \oplus E_{\lambda_2}(f) \oplus \cdots \oplus E_{\lambda_k}(f).$$

Proof. We need to show that if $v_i \in E_{\lambda_i}(f)$ for each $i \in \{1, 2, \dots, k\}$ are chosen such that $v_1 + v_2 + \cdots + v_k = 0$, then $v_1 = v_2 = \cdots = v_k = 0$. We will do so by induction over k.
In the base case of $k = 1$, the statement is obviously true. Assume nowas induction hypothesis that the sum $E_{\lambda_1}(f) \oplus \cdots \oplus E_{\lambda_{k-1}}(f)$ is direct. Since f is linear, if we apply f to both sides of

$$v_1 + v_2 + \cdots + v_k = 0, \tag{9.4}$$

we obtain

$$f(v_1) + f(v_2) + \cdots + f(v_k) = f(0) \quad \Leftrightarrow \quad \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k v_k = 0. \tag{9.5}$$

Taking the λ_k-fold multiple of (9.4) and subtracting it from (9.5) yields

$$(\lambda_1 - \lambda_k)v_1 + (\lambda_2 - \lambda_k)v_2 + \cdots + (\lambda_{k-1} - \lambda_k)v_{k-1} = 0. \tag{9.6}$$

Since eigenspaces are linear subspaces, $(\lambda_i - \lambda_k)v_i \in E_{\lambda_i}(f)$ for each $i \in \{1, 2, \dots, k-1\}$. By the induction hypothesis, we thus derive from (9.6) that $(\lambda_i - \lambda_k)v_i = 0$ for each $i \in \{1, 2, \dots, k-1\}$. Since by assumption $\lambda_i \neq \lambda_k$ for each $i \neq k$, we derive that $v_1 = v_2 = \cdots = v_{k-1} = 0$. Inserting this into (9.4) shows that $v_k = 0$ as well. This completes the induction step and thereby the proof. □

The previous lemma has an immediate consequence that is important to keep in mind.

Corollary 9.40. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. If $\lambda_1, \dots, \lambda_k \in K$ are pairwise distinct eigenvalues of f and if $v_i \in V$ is an eigenvector corresponding to λ_i for each $i \in \{1, 2, \dots, k\}$, then $\{v_1, v_2, \dots, v_k\}$ is linearly independent.*

Proof. Let $\mu_1, \dots, \mu_k \in K$, such that $\sum_{i=1}^{k} \mu_i v_i = 0$. Then $\mu_i v_i \in E_{\lambda_i}(f)$ for each $i \in \{1, 2, \dots, k\}$ and it follows from Lemma 9.39 that $\mu_i v_i = 0$ for each $i \in \{1, 2, \dots, k\}$. Since eigenvectors are by definition non-zero, this yields $\mu_1 = \mu_2 = \cdots = \mu_k = 0$, which shows that $\{v_1, \dots, v_k\}$ is linearly independent. □

Theorem 9.41. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. The following statements are equivalent:*

(i) f is diagonalizable.

(ii) There are pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_k \in K$ of f, such that

$$V = E_{\lambda_1}(f) \oplus E_{\lambda_2}(f) \oplus \cdots \oplus E_{\lambda_k}(f).$$

(iii) There are pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_k \in K$ of f, such that

$$\dim V = \dim E_{\lambda_1}(f) + \dim E_{\lambda_2}(f) + \cdots + \dim E_{\lambda_k}(f).$$

Proof. (ii) $\Rightarrow$ (iii): This follows immediately from Remark 9.38.
(iii) $\Rightarrow$ (i): Let $n_i := \dim E_{\lambda_i}(f)$ and B_i be a basis of $E_{\lambda_i}(f)$ for each $i \in \{1, 2, \ldots, k\}$. From the linear independence of the B_i and from Corollary 9.40 we derive that $B := B_1 \cup B_2 \cup \cdots \cup B_k$ is linearly independent. Since B has $n_1 + n_2 + \cdots + n_k = \dim V$ many elements, it follows from Corollary 6.33 that B is a basis of V. Hence, V possesses an eigenbasis for f.
(i) $\Rightarrow$ (ii): By Theorem/Definition 9.34, V has an eigenbasis B for f. Let $\lambda_1, \ldots, \lambda_k \in K$ be the pairwise distinct eigenvalues of f. For each $i \in \{1, 2, \ldots, k\}$ we let $B_i := B \cap E_{\lambda_i}(f)$. Using Corollary 6.33, it follows that B_i is a basis of $E_{\lambda_i}(f)$ for each $i \in \{1, 2, \ldots, k\}$. This shows that $V = E_{\lambda_1}(f) + \cdots + E_{\lambda_k}(f)$ and it follows from Lemma 9.39 that the sum is direct. □

This already gives us the diagonalizability of a map under a condition on the eigenvalues.

Corollary 9.42. *Let V be a finite-dimensional vector space over K, let $n = \dim V$ and let $f : V \to V$ be an endomorphism.*

a) f has at most n pairwise distinct eigenvalues.

b) If f has n pairwise distinct eigenvalues, then f is diagonalizable.

Proof. Let $\lambda_1, \ldots, \lambda_k \in K$ be pairwise distinct eigenvalues of f. For each i let $v_i \neq 0$ be an eigenvector corresponding to λ_i. Since $\operatorname{span}(\{v_i\}) \subset E_{\lambda_i}(f)$, it follows that $\dim E_{\lambda_i}(f) \geq 1$.

a) Using Lemma 9.39 and Remark 9.38.(2), we compute

$$k \leq \sum_{i=1}^{k} \dim E_{\lambda_i}(f) = \dim(E_{\lambda_1}(f) \oplus \cdots \oplus E_{\lambda_k}(f)) \leq \dim V = n.$$

Hence, $k \leq n$.

b) If $k = n$ in the situation of part a), then the chain of inequalities becomes a chain of equalities, in particular $\dim V = \sum_{i=1}^{k} \dim E_{\lambda_i}(f)$. By Theorem 9.41, this implies that f is diagonalizable.

□

So far, so good. Using Theorem 9.41 we can decide whether an endomorphism is diagonalizable by studying its eigenspaces. But to do so for a given endomorphism, we still need to find methods for finding the eigenvalues of an endomorphism. We are going to develop such a computational method using the determinant of an endomorphism and of a matrix.

Theorem 9.43. *Let V be a finite-dimensional vector space over K. let $f : V \to V$ be an endomorphism and let $\lambda \in K$. Then λ is an eigenvalue of K if and only if*

$$\det(f - \lambda \mathrm{id}_V) = 0.$$

Proof. First of all, we note that $f - \lambda \mathrm{id}_V$ is again an endomorphism. We derive from Theorem 9.21.b) and Corollary 7.23.c) that $\det(f - \lambda \mathrm{id}_V) = 0$ if and only if $f - \lambda \mathrm{id}_V$ is not surjective, i.e. if and only if $\operatorname{rank}(f - \lambda \mathrm{id}_V) < \dim V$. But by the dimension formula for linear maps (Theorem 7.22), this is in turn equivalent to

$$\dim \ker(f - \lambda \mathrm{id}_V) = \dim V - \operatorname{rank}(f - \lambda \mathrm{id}_V) > 0.$$

This holds true if and only $\ker(f - \lambda \mathrm{id}_V)$ contains some $v \neq 0$, which is equivalent to the existence of an eigenvector corresponding to λ. □

Since the determinant of an endomorphism is defined via its matrix representations, we will in the next section reformulate the condition that $\det(f - \lambda \mathrm{id}_V) = 0$ in terms of determinants of matrices.

9.5 Characteristic polynomials of endomorphisms and matrices

If we restrict the considerations from the previous section to maps of the form $f_A : K^n \to K^n$, where $A \in M(n \times n, K)$, we can associate eigenvalues and eigenvectors with matrices.

Definition 9.44. Let $n \in \mathbb{N}$ and let $A = (a_{ij}) \in M(n \times n, K)$.

a) We call $\lambda \in K$ an *eigenvalue of A* if λ is an eigenvalue of $f_A : K^n \to K^n$, i.e. if there exists $v \in K^n$ with $v \neq 0$, such that

$$Av = \lambda v.$$

v is then called an *eigenvector of A corresponding to λ* and $E_\lambda(A) := E_\lambda(f_A)$ is called the *eigenspace of A associated with λ*.

b) We call A *diagonalizable* if A is similar to a diagonal matrix.

c) The function

$$\chi_A : K \to K, \quad \chi_A(t) = \det(A - t \cdot I_n) = \det \begin{pmatrix} a_{11} - t & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - t & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - t \end{pmatrix},$$

is called *the characteristic polynomial of A*.

Remark 9.45. Let $n \in \mathbb{N}$ and $A \in M(n \times n, K)$.

(1) A is diagonalizable if and only if $f_A : K^n \to K^n$ is diagonalizable in the sense of Definition 9.33. This follows from Theorem 9.19 since as a matrix representation $A = M_B(f_A)$ for B the canonical basis of K^n.

(2) Considering the explicit form of the determinant (9.1), we see that the determinant is given as the sum of finitely many products of n many entries of the matrix. Thus, *the function $\chi_A(t)$ is indeed a polynomial in the variable t* whose coefficients are determined by the matrix entries.

(3) Taking a closer look at the formula (9.1) for the determinant and at the structure of $A - tI_n$ in Definition 9.44 one realizes that the only summand containing an n-fold product of t is the one belonging to $\sigma = \mathrm{id}$, which is given as

$$(a_{11} - t)(a_{22} - t) \dots (a_{nn} - t) = (-1)^n t^n + \dots .$$

Hence, χ_A is of the form

$$\chi_A(t) = (-1)^n t^n + c_{n-1} t^{n-1} + \dots + c_1 t + c_0$$

for suitable $c_0, c_1, \dots, c_{n-1} \in K$. (In fact, one can even show that $c_{n-1} = 0$.)

Corollary 9.46. *Let $n \in \mathbb{N}$, let $A \in M(n \times n, K)$ and let $\lambda \in K$. Then λ is an eigenvalue of A if and only if $\chi_A(t) = 0$.*

Proof. Apply Theorem 9.43 to $f_A : K^n \to K^n$. □

Example 9.47. Let $A = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \in M(2 \times 2, \mathbb{R})$. We compute its characteristic polynomial as

$$\chi_A(t) = \det \begin{pmatrix} -t & 1 \\ -2 & -3 - t \end{pmatrix} = (-t)(-3 - t) - 1 \cdot (-2) = t^2 + 3t + 2 = (t + 1)(t + 2).$$

Hence, $\chi_A(t) = 0$ if and only if $t \in \{-1, -2\}$, so the eigenvalues of A are $\lambda_1 = -1$ and $\lambda_2 = -2$. By Corollary 9.42, this already shows that A is diagonalizable.
To find eigenvectors of A corresponding to these eigenvalues, we have to solve systems of linear equations. Let $v = (x, y) \in \mathbb{R}^2$ be an eigenvector corresponding to -1. Then

$$\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = - \begin{pmatrix} x \\ y \end{pmatrix} \quad \Leftrightarrow \quad \begin{cases} y & = -x, \\ -2x - 3y & = -y. \end{cases} \quad \Leftrightarrow \quad x + y = 0.$$

In particular, $(1, -1)$ is an eigenvector of A corresponding to -1. Analogously, one shows that $(1, -2)$ is an eigenvector of A corresponding to -2. Hence, $\mathbb{R}^2 = E_{-1}(A) \oplus E_{-2}(A)$, where $E_{-1}(A) = \{(x, -x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ and $E_{-2}(A) = \{(x, -2x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$.

Theorem 9.48. *Let $n \in \mathbb{N}$ and let $A, B \in M(n \times n, K)$. If A and B are similar, then $\chi_A = \chi_B$.*

Proof. Let $S \in \mathrm{GL}(n, K)$, such that $B = SAS^{-1}$, and let $t \in K$. Then, by Theorem 9.13,

$$B - tI_n = SAS^{-1} - tS \cdot I_n \cdot S^{-1} = S(AS^{-1} - tI_n S^{-1}) = S(A - tI_n)S^{-1}.$$

Hence, by Corollary 9.15.b),

$$\chi_B(t) = \det(B - tI_n) = \det(S(A - tI_n)S^{-1}) = \det(A - tI_n) = \chi_A(t).$$

□

We now transfer the notion of characteristic polynomials from matrices to endomorphisms.

Definition 9.49. Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. The *characteristic polynomial of f* is given by

$$\chi_f : K \to K, \qquad \chi_f(t) = \det(f - t \cdot \mathrm{id}_V).$$

Corollary 9.50. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. Let B be an ordered basis of V. Then*

$$\chi_f = \chi_{M_B(f)}.$$

Proof. This follows immediately from combining Theorem 9.19 with Theorem 9.48. □

In the language of characteristic polynomials, we can rephrase Theorem 9.43 for an endomorphism $f : V \to V$ as follows:

$$\lambda \in K \text{ is an eigenvalue of } f \quad \Leftrightarrow \quad \chi_f(\lambda) = 0.$$

We want to take another look at diagonalizability using characteristic polynomials.

Theorem 9.51. *Let $n \in \mathbb{N}$ and V be an n-dimensional vector space over K and let $f : V \to V$ be an endomorphism. If f is diagonalizable, then there are eigenvalues $\lambda_1, \dots, \lambda_n \in K$ of f, such that*

$$\chi_f(t) = (\lambda_1 - t)(\lambda_2 - t) \dots (\lambda_n - t) \qquad \forall t \in K.$$

Proof. Let B be a basis of V, such that

$$A := M_B(f) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \in M(n \times n, K).$$

Then by Corollary 9.50

$$\chi_f(t) = \chi_A(t) = \det \begin{pmatrix} \lambda_1 - t & 0 & \dots & 0 \\ 0 & \lambda_2 - t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n - t \end{pmatrix} = (\lambda_1 - t)(\lambda_2 - t) \dots (\lambda_n - t).$$

□

Remark 9.52. (1) Note that in Theorem 9.51, the eigenvalues are not necessarily pairwise distinct, i.e. the characteristic polynomial is in general of the form

$$\chi_f(t) = (-1)^n (t - \lambda_1)^{n_1} (t - \lambda_2)^{n_2} \dots (t - \lambda_k)^{n_k},$$

where $\lambda_1, \dots, \lambda_k \in K$ are pairwise distinct and where $n_1, \dots, n_k \in \mathbb{N}$.

(2) The converse statement of Theorem 9.51 does not hold. Consider the map $f_A : \mathbb{R}^2 \to \mathbb{R}^2$ induced by $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. One checks that $\chi_A(t) = (1-t)^2 = (t-1)(t-1)$, so A has only 1 as eigenvalue. So if f_A was diagonalizable, then $V = E_1(A)$ by Theorem 9.41. But this is false, since for example $f_A(0,1) = (1,1)$.

Example 9.53. (1) Consider $A = \begin{pmatrix} 1 & 2 & 3 \\ -1 & -1 & 4 \\ 0 & 0 & 5 \end{pmatrix} \in M(3 \times 3, \mathbb{R})$. Then by Exercise 3.a) from Sheet 5, we compute

$$\begin{aligned} \chi_A(t) &= \det \begin{pmatrix} 1-t & 2 & 3 \\ -1 & -1-t & 4 \\ 0 & 0 & 5-t \end{pmatrix} = \det \begin{pmatrix} 1-t & 2 \\ -1 & -1-t \end{pmatrix} \cdot (5-t) \\ &= ((1-t)(-1-t)+2)(5-t) = (t^2+1)(5-t). \end{aligned}$$

Since $t^2 + 1$ does not have any solutions in $\mathbb{R}$, we cannot decompose χ_A into linear factors, hence $f_A : \mathbb{R}^3 \to \mathbb{R}^3$ and thus A is not diagonalizable by Theorem 9.51.

(2) We consider the same A as before, but this time view it as a *complex* matrix $A \in M(3 \times 3, \mathbb{C})$. Note that in the same way as for the real numbers, A induces a matrix $f_A : \mathbb{C}^3 \to \mathbb{C}^3$ between vector spaces over $\mathbb{C}$. Since we now view χ_A as a *complex* polynomial, we can compute

$$\chi_A(t) = (t^2 + 1)(5 - t) = -(t - i)(t + i)(t - 5).$$

Hence, $f_A : \mathbb{C}^3 \to \mathbb{C}^3$ has the three eigenvalues 5, i and $-i$, so by Corollary 9.42 the endomorphism f_A is diagonalizable.

We have reduced the search for eigenvalues of endomorphisms to the search for zeros of polynomials. This is an important and well-studied mathematical problem of its own and due to the limited time in this course, we will not go into detail about this problem.[3] Instead, we will study some properties of characteristic polynomials that allow for another characterization of diagonalizability.

The property of a polynomial of being a product of linear terms has a particular name in the theory of polynomials.

Definition 9.54. Let $n \in \mathbb{N}$ and let $p : K \to K$ be a polynomial of degree n. We say that p *decomposes into linear factors* if there are $a_1, \dots, a_n \in K$ and $c \in K \setminus \{0\}$, such that

$$p(x) = c(x - a_1)(x - a_2) \dots (x - a_n) \qquad \forall x \in K.$$

The importance of such linear factors stems from the following result, which we shall not prove in this course, since it would require a more thorough study of properties of polynomials.

Theorem 9.55. *Let $p : K \to K$ be a polynomial. If $a \in K$ satisfies $p(a) = 0$, then there exists a polynomial $q : K \to K$ with $\deg q = \deg p - 1$, such that*

$$p(x) = (x - a)q(x) \qquad \forall x \in K.$$

One can applying this result iteratively, by "factoring out" one linear factor and applying Theorem 9.55 to the remaining factor over and over. Eventually, one ends up with the folloiwng statement.

Corollary 9.56. *Let $p : K \to K$ be a polynomial of degree $n \in \mathbb{N}$. There are $a_1, \dots, a_r \in K$, $k_1, \dots, k_r \in \mathbb{N}$ and a polynomial $q : K \to K$ with $q(x) \neq 0$ for all $x \in K$ and $k_1 + \dots + k_r + \deg q = n$, such that*

$$p(x) = (x - a_1)^{k_1} \dots (x - a_r)^{k_r} q(x) \qquad \forall x \in K.$$

Remark 9.57. *Every complex polynomial $p : \mathbb{C} \to \mathbb{C}$ decomposes into linear factors.* This follows from Corollary 9.56 and the fundamental theorem of algebra (Theorem 1.75), since the only complex polynomials that vanish nowhere in $\mathbb{C}$ are the (non-zero) constant ones.

Definition 9.58. Let V be a finite-dimensional vector space over K, let $f : V \to V$ be an endomorphism and let $\lambda \in K$ be an eigenvalue of f.

a) We call $m_{\text{geom}}(\lambda, f) := \dim E_\lambda(f)$ the *geometric multiplicity of λ*.

[3] If you are interested in further details, just take your favorite linear algebra textbook and look up minimal polynomials and the Cayley-Hamilton theorem.

b) Let $\chi_f : K \to K$ be the characteristic polynomial. We call the number $m \in \mathbb{N}$ for which there exists a polynomial $p : K \to K$ with $p(\lambda) \neq 0$, such that

$$\chi_f(t) = (t - \lambda)^k p(t) \quad \forall t \in K$$

the *algebraic multiplicity of* λ and denote it by $m_{\text{alg}}(\lambda, f) := m$.

Lemma 9.59. *Let V be a finite-dimensional vector space over K, let $f : V \to V$ be an endomorphism and let $\lambda \in K$ be an eigenvalue of f. Then*

$$m_{geom}(\lambda, f) \leq m_{alg}(\lambda, f).$$

Proof. Put $n := \dim V$, $r := m_{\text{geom}}(\lambda, f)$ and let $(b_1, \ldots, b_r)$ be an ordered basis of $E_\lambda(f)$. By the basis extension theorem, we can extend this to an ordered basis $B = (b_1, \ldots, b_r, b_{r+1}, \ldots, b_n)$. By definition of eigenvectors, it is evident that the matrix representation $M_B(f)$ is as a block matrix of the form

$$M_B(f) = \begin{pmatrix} \lambda I_r & A \\ 0 & C \end{pmatrix}, \qquad \text{where } A \in M(r \times (n-r), K),\ C \in M((n-r) \times (n-r), K).$$

By Exercise 3a) from Sheet 5, it holds that

$$\begin{aligned} \chi_f(t) = \chi_{M_B(f)}(t) &= \det(M_B(f) - tI_n) = \det \begin{pmatrix} (\lambda - t) I_r & A \\ 0 & C - tI_{n-r} \end{pmatrix} \\ &= \det((\lambda - t) I_r) \cdot \det(C - tI_{n-r}) \\ &= (\lambda - t)^r \chi_C(t) = (-1)^r (t - \lambda)^r \chi_C(t). \end{aligned}$$

Since $\chi_C(t)$ is again a polynomial (which might contain a factor of the form $(t - \lambda)^k$ or not), this shows that $m_{\text{alg}}(\lambda, f) \geq r$, which we wanted to show. □

Theorem 9.60. *Let V be a finite-dimensional vector space over K and let $f : V \to V$ be an endomorphism. The following statements are equivalent:*

(i) f is diagonalizable.

(ii) χ_f decomposes into linear factors and for every eigenvalue λ of f it holds that

$$m_{geom}(\lambda, f) = m_{alg}(\lambda, f).$$

Proof. Let $\lambda_1, \ldots, \lambda_k \in K$ be the eigenvalues of f, given such that $\lambda_1, \ldots, \lambda_k$ are pairwise distinct. Put $n := \dim V$.
(i) $\Rightarrow$ (ii): We have already seen in Theorem 9.51 that χ_f decomposes into linear factors if f is diagonalizable. With $n_i := m_{\text{alg}}(f, \lambda)$ for each $i \in \{1, 2, \ldots, k\}$ we obtain $\chi_f(t) = (-1)^n (t - \lambda_1)^{n_1} \ldots (t - \lambda_k)^{n_k}$ for all $t \in K$. Since $\deg \chi_f = n$, it follows that

$$\sum_{i=1}^{k} m_{\text{alg}}(\lambda_i, f) = \sum_{i=1}^{k} n_i = n.$$

Since f is diagonalizable, it further holds by Theorem 9.41 that

$$\sum_{i=1}^{k} m_{\text{geom}}(\lambda_i, f) = \sum_{i=1}^{k} \dim E_{\lambda_i}(f) = \dim V = n \quad \Rightarrow \quad \sum_{i=1}^{k} m_{\text{geom}}(\lambda_i, f) = \sum_{i=1}^{k} m_{\text{alg}}(\lambda_i, f).$$

Since $m_{\text{geom}}(\lambda_i, f) \leq m_{\text{alg}}(\lambda_i, f)$ for each i by Lemma 9.59, this implies $m_{\text{geom}}(\lambda_i, f) = m_{\text{alg}}(\lambda_i, f)$ for each $i \in \{1, 2, \ldots, k\}$.

(ii) $\Rightarrow$ (i): Since χ_f decomposes into linear factors, it holds as in the other implication that $\sum_{i=1}^{k} m_{\text{alg}}(\lambda_i, f) = n$. By assumption, this shows that

$$\sum_{i=1}^{k} \dim E_{\lambda_i}(f) = \sum_{i=1}^{k} m_{\text{geom}}(\lambda_i, f) = \sum_{i=1}^{k} m_{\text{alg}}(\lambda_i, f) = n = \dim V.$$

Thus, Theorem 9.41 implies that f is diagonalizable. □

The results of this section, especially Theorem 9.60, give us a method at hand for deciding whether a given endomorphism is diagonalizable or not and, in case of diagonalizability, to find a basis for which the matrix representation is a diagonal matrix.

Method for the diagonalization of an endomorphism $f : V \to V$

(1) Compute its characteristic polynomial χ_f and check if it decomposes into linear factors over K. If not, then f is not diagonalizable.

(2) If χ_f decomposes into linear factors, read off the eigenvalues and their algebraic multiplicities from that decomposition.

(3) Compute the geometric multiplicity of each eigenvalue λ of f by finding the dimension of the solution set of $f - \lambda \text{id}_V = 0$. By coordinate representations, this amounts to solving a system of linear equations e.g. by Gaussian elimination.

(4) If the geometric and algebraic multiplicities coincide for each eigenvalue, then f is diagonalizable. Otherwise, f is not diagonalizable.

(5) If f is diagonalizable, use the results of step (3) to find an ordered base of the corresponding eigenspace for each eigenvalue of f. Put these bases together to obtain a basis B of V for which $M_B(f)$ is a diagonal matrix.

Chapter 10

Euclidean and unitary vector spaces

10.1 Inner products on real and complex vector spaces

We want to introduce some additional structures on real and complex[1] vector spaces which have a geometric meaning and which are used to define notions of *length* or *distance* of vectors. We will also introduce *angles* in real vector spaces by means of these structures.

Definition 10.1. Let V be a *real* vector space. An *inner product* or *scalar product on* V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

which has the following properties:

(i) (bilinearity) For all $v_1, v_2, v, w \in V$ and $\lambda \in \mathbb{R}$ it holds that

$$\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle \quad \text{and} \quad \langle \lambda v, w \rangle = \lambda \langle v, w \rangle.$$

(In other words, the map $\langle \cdot, w \rangle : V \to \mathbb{R}$ is linear for each $w \in V$.)

(ii) (symmetry) $\langle v, w \rangle = \langle w, v \rangle$ for all $v, w \in V$.

(iii) (positive definiteness) $\langle v, v \rangle > 0$ for all $v \in V \setminus \{0\}$.

If $\langle \cdot, \cdot \rangle$ is an inner product on V, then the pair $(V, \langle \cdot, \cdot \rangle)$ is called a *Euclidean vector space*. If it is apparent which inner product we are using, we simply write V instead of $(V, \langle \cdot, \cdot \rangle)$.

Remark 10.2. Let $(V, \langle \cdot, \cdot \rangle)$ be a Euclidean vector space.

(1) For all $v, w, w_1, w_2 \in V$ we compute that

$$\langle v, w_1 + w_2 \rangle \overset{\text{(ii)}}{=} \langle w_1 + w_2, v \rangle \overset{\text{(i)}}{=} \langle w_1, v \rangle + \langle w_2, v \rangle \overset{\text{(i)}}{=} \langle v, w_1 \rangle + \langle v, w_2 \rangle,$$
$$\langle v, \lambda w \rangle \overset{\text{(ii)}}{=} \langle \lambda w, v \rangle \overset{\text{(i)}}{=} \lambda \langle w, v \rangle \overset{\text{(ii)}}{=} \lambda \langle v, w \rangle.$$

This shows that the map $\langle v, \cdot \rangle : V \to \mathbb{R}$ is linear for all $v \in V$.

(2) Properties (i) and (ii) of an inner product yield $\langle v, 0 \rangle = \langle 0, v \rangle = 0$ for all $v \in V$.

[1] Good news for a reader who feel uneasy working with abstract fields: in this whole chapter, we will only consider the fields $K = \mathbb{R}$ and $K = \mathbb{C}$.

Example 10.3. (1) Let $n \in \mathbb{N}$. An inner product on $\mathbb{R}^n$ is given by

$$\langle (x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

for all $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$. This inner product is called the *standard inner product* or *Euclidean inner product* on $\mathbb{R}^n$.

(2) If $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the standard inner product, then for each $A \in \mathrm{GL}(n, K)$ the map $\langle \cdot, \cdot \rangle_A : \mathbb{R}^n \times \mathbb{R}^n$ given by

$$\langle v, w \rangle_A := \langle Av, Aw \rangle$$

is an inner product on $\mathbb{R}^n$. (Later we will see, that *every* inner product on $\mathbb{R}^n$ is given in this way.)

(3) Let $a, b \in \mathbb{R}$ with $a < b$. Then we can define an inner product on the real vector space $C^0([a, b], \mathbb{R})$ by

$$\langle f, g \rangle := \int_a^b f(x) g(x)\, dx \qquad \forall f, g \in C^0([a, b], \mathbb{R}).$$

Its bilinearity property follows from the properties of Riemann integrals, while the symmetry property is evident. Concerning positive definiteness: Since $f(x)^2 \geq 0$ for all $x \in [a, b]$ and $f \in C^0([a, b], \mathbb{R})$, the monotonicity property of Riemann integrals yields $\langle f, f \rangle \geq 0$ for all $f \in C^0([a, b], \mathbb{R})$. However, if $0 = \langle f, f \rangle = \int_a^b f(x)^2\, dx$, then it follows from Exercise 1 of Sheet 11 of Mathematics for Physicists 1 that $f = 0$. Thus, if $f \neq 0$, then $\langle f, f \rangle > 0$, which shows property (iii) of an inner product.

Definition 10.4. Let V be a *complex* vector space. An *inner product on* V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$$

which has the following properties:

(i) For all $v_1, v_2, v, w \in V$ and $\lambda \in \mathbb{C}$ it holds that

$$\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle \quad \text{and} \quad \langle \lambda v, w \rangle = \lambda \langle v, w \rangle.$$

(In other words, the map $\langle \cdot, w \rangle : V \to \mathbb{C}$ is linear for each $w \in V$.)

(ii) $\langle v, w \rangle = \overline{\langle w, v \rangle}$ for all $v, w \in V$. (Here, the bar denotes complex conjugation.)

(iii) $\langle v, v \rangle \in (0, +\infty)$ for all $v \in V \setminus \{0\}$.

If $\langle \cdot, \cdot \rangle$ is an inner product on V, then the pair $(V, \langle \cdot, \cdot \rangle)$ is called a *unitary vector space*. If it is apparent which inner product we are using, we simply write V instead of $(V, \langle \cdot, \cdot \rangle)$.

Remark 10.5. Let $(V, \langle \cdot, \cdot \rangle)$ be a unitary vector space.

(1) For all $v, w, w_1, w_2 \in V$ we compute that

$$\langle v, w_1 + w_2 \rangle \overset{\text{(ii)}}{=} \overline{\langle w_1 + w_2, v \rangle} \overset{\text{(i)}}{=} \overline{\langle w_1, v \rangle + \langle w_2, v \rangle} = \overline{\langle w_1, v \rangle} + \overline{\langle w_2, v \rangle} \overset{\text{(i)}}{=} \langle v, w_1 \rangle + \langle v, w_2 \rangle,$$
$$\langle v, \lambda w \rangle \overset{\text{(ii)}}{=} \overline{\langle \lambda w, v \rangle} \overset{\text{(i)}}{=} \overline{\lambda \langle w, v \rangle} = \overline{\lambda} \cdot \overline{\langle w, v \rangle} \overset{\text{(ii)}}{=} \overline{\lambda} \langle v, w \rangle.$$

This property of the map $\langle v, \cdot \rangle : V \to \mathbb{C}$ is called *semilinearity* or *complex antilinearity*.

(2) As for Euclidean vector spaces, it holds that $\langle v, 0\rangle = \langle 0, v\rangle = 0$ for all $v \in V$.

Example 10.6. Let $n \in \mathbb{N}$. An inner product on $\mathbb{C}^n$ is given by

$$\langle (z_1, z_2, \ldots, z_n), (w_1, w_2, \ldots, w_n)\rangle = z_1\overline{w}_1 + z_2\overline{w}_2 + \cdots + z_n\overline{w}_n$$

for all $(z_1, z_2, \ldots, z_n), (w_1, w_2, \ldots, w_n) \in \mathbb{C}^n$.

Theorem 10.7 (Cauchy-Schwarz inequality)**.** *Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean or a unitary vector space. Then*

$$|\langle v, w\rangle|^2 \leq \langle v, v\rangle \cdot \langle w, w\rangle \qquad \forall v, w \in V,$$

where equality holds if and only if v and w are linearly dependent.

Proof. We will only show the unitary case as the Euclidean case is shown completely analogously. For $w = 0$ both sides of the inequality vanish by Remark 10.5.(2), so the claim holds true in this case. Assume throughout the following that $w \neq 0$, such that $\langle w, w\rangle > 0$. We compute for all $\lambda \in \mathbb{C}$:

$$0 \overset{\text{(iii)}}{\leq} \langle v - \lambda w, v - \lambda w\rangle \overset{\text{(i)}}{=} \langle v, w - \lambda w\rangle + \langle -\lambda w, v - \lambda w\rangle .$$

Using properties (i) and (ii) of an inner product as well as Remark 10.5.(1), we derive

$$\begin{aligned} 0 &\leq \langle v, v\rangle + \langle v, -\lambda w\rangle + \langle -\lambda w, v\rangle + \langle -\lambda w, -\lambda, w\rangle \\ &= \langle v, v\rangle - \overline{\lambda}\,\langle v, w\rangle - \lambda\overline{\langle v, w\rangle} + \lambda\overline{\lambda}\,\langle w, w\rangle . \end{aligned}$$

Inserting $\lambda = \frac{\langle v,w\rangle}{\langle w,w\rangle}$ yields

$$\begin{aligned} & 0 \leq \langle v, v\rangle - \frac{\overline{\langle v, w\rangle}}{\langle w, w\rangle}\langle v, w\rangle - \frac{\langle v, w\rangle}{\langle w, w\rangle}\overline{\langle v, w\rangle} + \frac{\langle v, w\rangle \cdot \overline{\langle v, w\rangle}}{\langle w, w\rangle \cdot \langle w, w\rangle}\langle w, w\rangle \\ \Leftrightarrow \quad & 0 \leq \langle v, v\rangle - \frac{|\langle v, w\rangle|^2}{\langle w, w\rangle}. \end{aligned}$$

Reordering the terms in the last inequality shows the claimed inequality. Moreover, one sees from the proof that equality holds if and only if $0 = \langle v - \lambda w, v - \lambda w\rangle$ which is in turn equivalent to $v = \lambda w$, i.e. v and w are linearly dependent. □

Definition 10.8. Let V a vector space over $K = \mathbb{R}$ or $K = \mathbb{C}$. A *norm on V* is a map $\|\cdot\| : V \to \mathbb{R}$ which has the following properties:

(i) $\|v\| \geq 0$ for all $v \in V$ and $\|v\| = 0$ holds if and only if $v = 0$.

(ii) $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all $\lambda \in \mathbb{C}$ and $v \in V$.

(iii) (triangle inequality) $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.

The pair $(V, \|\cdot\|)$ is called a *normed vector space*. If it is apparent which norm we are using, we simply write V instead of $(V, \|\cdot\|)$.

The connections between inner products and norms is given by the following observation.

Theorem/Definition 10.9. *Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean or a unitary vector space. Then $\|\cdot\| : V \to \mathbb{R}$ is a norm on V, where*

$$\|v\| := \sqrt{\langle v, v\rangle} \quad \forall v \in V.$$

We call $\|\cdot\|$ *the norm induced by* $\langle\cdot,\cdot\rangle$.

Proof. By property (iii) of a real or complex inner product, the term inside the square root is always non-negative, so $\|\cdot\|$ is indeed well-defined. We need to show that $\|\cdot\|$ satisfies the three properties of a norm. We again consider only the complex case as the real case is proven along the same lines.

(i) It follows from property (iii) of an inner product and the non-negativity of roots that $\|v\| \geq 0$ for all $v \in V$. Moreover, $0 = \|v\| = \sqrt{\langle v, v\rangle}$ implies $\langle v, v\rangle = 0$, which implies $v = 0$ by property (iii) of an inner product.

(ii) For $\lambda \in \mathbb{C}$ and $v \in V$ we compute from property (i) of an inner product and Remark 10.5.(1) that

$$\|\lambda v\| = \sqrt{\langle \lambda v, \lambda v\rangle} = \sqrt{\lambda \langle v, \lambda v\rangle} = \sqrt{\lambda\overline{\lambda}\langle v, v\rangle} = \sqrt{|\lambda|^2 \langle v, v\rangle} = |\lambda| \cdot \|v\|.$$

(iii) For $v, w \in V$ we compute

$$\begin{aligned}\|v + w\|^2 &= \langle v + w, v + w\rangle = \langle v, v\rangle + \langle v, w\rangle + \langle w, v\rangle + \langle w, w\rangle \\ &= \|v\|^2 + \langle v, w\rangle + \overline{\langle v, w\rangle} + \|w\|^2.\end{aligned}$$

By Theorem 1.70.b) and Theorem 1.73.c), this yields

$$\begin{aligned}\|v + w\|^2 &= \|v\|^2 + 2\mathrm{Re}(\langle v, w\rangle) + \|w\|^2 \\ &\leq \|v\|^2 + 2\,|\langle v, w\rangle| + \|w\|^2 \\ &\overset{\text{Theorem 10.7}}{\leq} \|v\|^2 + 2\sqrt{\langle v, v\rangle \cdot \langle w, w\rangle} + \|w\|^2 \\ &= \|v\|^2 + 2\|v\| \cdot \|w\| + \|w\|^2 = (\|v\| + \|w\|)^2.\end{aligned}$$

Taking square roots shows that property (iii) of a norm is satisfied.

□

Remark 10.10. Note that in terms of the norm induced by an inner product in a Euclidean or a unitary vector space $(V, \langle\cdot,\cdot\rangle)$, the Cauchy-Schwarz inequality is equivalent to

$$|\langle v, w\rangle| \leq \|v\| \cdot \|w\| \quad \forall v, w \in V. \tag{10.1}$$

Example 10.11. (1) The norm induced by the standard inner product on $\mathbb{R}^n$, explicitly

$$\|(x_1, x_2, \ldots, x_n)\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \qquad \forall (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n,$$

is called the *Euclidean norm on* $\mathbb{R}^n$.

(2) The norm induced by the standard inner product on $\mathbb{C}^n$, explicitly

$$\|(z_1, z_2, \ldots, z_n)\| = \sqrt{z_1\overline{z}_1 + z_2\overline{z}_2 + \cdots + z_n\overline{z}_n} = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2}$$

for all $(z_1, z_2, \ldots, z_n) \in \mathbb{C}^n$, is called the *Euclidean norm on* $\mathbb{C}^n$.

In the following, we will consider every Euclidean or unitary vector spaces as equipped with the norm induced by its inner product without explicitly mentioning this.

Theorem 10.12. *Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean vector space. Then for all $v, w \in V$ the following identities hold:*

a) (polarization identity) $\langle v, w\rangle = \frac{1}{2}(\|v+w\|^2 - \|v\|^2 - \|w\|^2)$.

b) (parallelogram law[2]) $\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2$

Proof. (This proof is not discussed in the lecture videos.)

a) We compute from the properties of real inner products that

$$\begin{aligned}\|v+w\|^2 &= \langle v+w, v+2w\rangle = \langle v,v\rangle + \langle v,w\rangle + \langle w,v\rangle + \langle w,w\rangle \\ &= \|v\|^2 + \|w\|^2 + 2\langle v,w\rangle .\end{aligned}$$

This shows the claim after rearranging the terms.

b) We compute that

$$\begin{aligned}\|v+w\|^2 &= \langle v,v\rangle + \langle v,w\rangle + \langle w,v\rangle + \langle w,w\rangle = \langle v,v\rangle + 2\langle v,w\rangle + \langle w,w\rangle ,\\ \|v-w\|^2 &= \langle v,v\rangle - \langle v,w\rangle - \langle w,v\rangle + \langle w,w\rangle = \langle v,v\rangle - 2\langle v,w\rangle + \langle w,w\rangle .\end{aligned}$$

Adding the two expressions yields

$$\|v+w\|^2 + \|v-w\|^2 = 2\langle v,v\rangle + 2\langle w,w\rangle = 2\|v\|^2 + 2\|w\|^2.$$

□

A formula that we shall omit and that is similar to the polarization identity holds in unitary vector spaces as well, i.e. inner products on complex vector spaces can be computed purely in terms of the norms they induce.

Theorem/Definition 10.13. *Let X be a set, let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let*

$$\mathcal{B}(X,K) := \{f : X \to K \mid f \text{ is bounded}\},$$

which is a linear subspace of K^X, hence a vector space over K. Then

$$\|\cdot\|_\infty : \mathcal{B}(X,K) \to \mathbb{R}, \qquad \|f\|_\infty = \sup_{x\in X} |f(x)|,$$

is a norm on $\mathcal{B}(X,K)$. It is called uniform norm *or* sup norm.

Proof. We need to check the three properties of a norm.

(i) Evidently, $\|f\|_\infty \geq 0$ for all $f \in \mathcal{B}(X,K)$. Moreover, we compute that

$$\|f\|_\infty = 0 \quad\Leftrightarrow\quad \sup_{x\in B} \underbrace{|f(x)|}_{\geq 0} = 0 \quad\Leftrightarrow\quad |f(x)| = 0 \ \forall x \in B \quad\Leftrightarrow\quad f = 0.$$

[2]Please consider the corresponding lecture video for a geometric explanation of that name.

(ii) By the properties of suprema that we studied in Mathematics 1, we compute for all $\lambda \in K$ and $f \in \mathcal{B}(X, K)$:

$$\|\lambda f\|_\infty = \sup_{x \in B} |\lambda f(x)| = \sup_{x \in B} (|\lambda| \cdot |f(x)|) = |\lambda| \sup_{x \in B} |f(x)| = |\lambda| \cdot \|f\|_\infty.$$

(iii) By the properties of suprema, it holds that $\sup(A_1 + A_2) \leq \sup A_1 + \sup A_2$ for all $A_1, A_2 \subset \mathbb{R}$. From this observation and the triangle inequality of the absolute value, we derive

$$\begin{aligned} \|f + g\|_\infty &= \sup_{x \in B} |f(x) + g(x)| \leq \sup_{x \in B} (|f(x)| + |g(x)|) \\ &\leq \sup_{x \in B} |f(x)| + \sup_{x \in B} |g(x)| = \|f\|_\infty + \|g\|_\infty. \end{aligned}$$

□

Example 10.14. We want to use the sup norm to construct an example of a norm that is *not* induced by an inner product, namely the sup norm on $\mathcal{B}([0, 2\pi], \mathbb{R})$. Put $f, g : [0, 2\pi] \to \mathbb{R}$, $f := \max(\sin, 0)$, $g := \max(-\sin, 0)$. One checks that

$$f(x) - g(x) = \sin x \quad \text{and} \quad f(x) + g(x) = |\sin x| \quad \forall x \in \mathbb{R}.$$

From this observation and the definition of the sup norm, we derive that

$$\|f\|_\infty = 1, \quad \|g\|_\infty = 1, \quad \|f + g\|_\infty = 1, \quad \|f - g\| = 1.$$

Hence,

$$\|f + g\|_\infty^2 + \|f - g\|_\infty^2 = 2 \neq 4 = 2\|f\|_\infty^2 + 2\|g\|_\infty^2,$$

so $\|\cdot\|_\infty$ violates the parallelogram law. Hence, $\|\cdot\|_\infty$ cannot be induced by an inner product.

Let $(V, \langle \cdot, \cdot \rangle)$ be a Euclidean vector space. By (10.1), it holds for all $v, w \in V \setminus \{0\}$ that

$$-1 \leq \frac{\langle v, w \rangle}{\|v\| \|w\|} \leq 1.$$

This shows that the term introduced in the following definition is well-defined:

Definition 10.15. Let $(V, \langle \cdot, \cdot \rangle)$ be a Euclidean vector space and let $v, w \in V \setminus \{0\}$. The *angle between v and w* is given by

$$\angle(v, w) \in [0, \pi], \qquad \angle(v, w) := \arccos\left(\frac{\langle v, w \rangle}{\|v\| \|w\|}\right).$$

Remark 10.16. Let $(V, \langle \cdot, \cdot \rangle)$ be Euclidean vector space. By the symmetry property of inner products, $\angle(v, w) = \angle(w, v)$.

Using this definition of an angle between two vectors and interpreting the norm of a vector as its length, we can recover a theorem from high school geometry by our new methods.

Theorem 10.17 (Law of cosines)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be Euclidean vector space and let $v, w \in V \setminus \{0\}$. Then*

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\|v\| \|w\| \cos(\angle(v, w)).$$

In particular, if $\angle(v, w) = \frac{\pi}{2}$ (i.e. if $\langle v, w \rangle = 0$), then the Pythagorean theorem holds:

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2.$$

Proof. We compute using the properties of real inner products:

$$\begin{aligned}\|v-w\|^2 &= \langle v-w, v-w\rangle = \langle v,w\rangle - \langle v,w\rangle - \langle w,v\rangle + \langle w,w\rangle \\ &= \|v\|^2 + \|w\|^2 - 2\langle v,w\rangle \\ &= \|v\|^2 + \|w\|^2 - 2\|v\|\|w\| \cdot \frac{\langle v,w\rangle}{\|v\|\|w\|} \\ &= \|v\|^2 + \|w\|^2 - 2\|v\|\|w\|\cos(\angle(v,w)).\end{aligned}$$

□

10.2 Orthogonal vectors and the Gram-Schmidt process

In K^n we prefer to always work with the canonical basis because of its simplicity and its useful features. In this section, we want to model important properties of the canonical basis and generalize some of its properties to arbitrary Euclidean and unitary vector spaces. A particular feature of the canonical basis is that each of its vectors has the same length, i.e. the same Euclidean norm. If we would for example replace (e_1, e_2, e_3) by the basis $B = (10000e_1, e_2, e_3)$ of $\mathbb{R}^3$, then the vector $(1,1,1)$ would have the coordinates $(\frac{1}{10000}, 1, 1)$ with respect to B, which is very counter-intuitive. Hence, one requirement for a basis resembling the canonical basis is that all of its vectors should have the same norm.
Another important feature of the canonical basis of $\mathbb{R}^2$ or $\mathbb{R}^3$ is that in our intuitive understanding any two distinct vectors from this basis are orthogonal or perpendicular to each other. In terms of the standard inner products on $\mathbb{R}^2$ and $\mathbb{R}^3$ one checks that two vectors v and w are orthogonal to each other if and only if $\langle v,w\rangle = 0$. This now indicates how we can generalize the notion of orthogonality to arbitrary Euclidean and unitary vector spaces. We want the vectors in a basis resembling the canonical basis to be pairwise orthogonal to each other as well.
A basis having these norm and orthogonality properties will be called an orthonormal basis and as we will see, such a basis has several computational advantanges over arbitrary ones. Without further ado, we next introduce and formalize the notions of orthogonality that we just lined out.

Definition 10.18. Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean or a unitary vector space.

a) Two vectors $u, v \in V$ are *orthogonal* if $\langle u, v\rangle = 0$. If u and v are orthogonal, one also writes $u \perp v$.

b) Let $M_1, M_2 \subset V$. We say that M_1 and M_2 are orthogonal and write $M_1 \perp M_2$ if $v_1 \perp v_2$ for all $v_1 \in M_1$ and $v_2 \in M_2$. If $M_1 = \{v\}$ consists of one single vector, then we write $v \perp M_2$ instead of $\{v\} \perp M_2$.

c) Let I be a set. A family $\{v_i \in V \mid i \in I\}$ is *orthogonal* if and $v_i \perp v_j$ for all $i, j \in I$ with $i \neq j$.

d) An orthogonal family $\{v_i \in V \mid i \in I\}$ is *orthonormal* if $\langle v_i, v_i\rangle = 1$ for each $i \in I$.

e) Let $(V, \langle\cdot,\cdot\rangle)$ be a finite-dimensional Euclidean or unitary vector space. A basis B of V is called an *orthonormal basis of V* if B is an orthonormal family.

Example 10.19. Let $K = \mathbb{R}$ or $K = \mathbb{C}$

(1) Let $(2,1), (1,-2) \in K^2$ and consider the standard inner product on K^2. Then

$$\langle (2,1), (1,-2) \rangle = 2 \cdot 1 + 1 \cdot (-2) = 0,$$

hence $(2,1)$ and $(1,-2)$ are orthogonal, However, $\{(2,1), (1,-2)\}$ is not orthonormal, as

$$\|(2,1)\| = \sqrt{2^2 + 1^2} = \sqrt{5}, \qquad \|(1,-2)\| = \sqrt{5}.$$

But this yields that $\{\frac{1}{\sqrt{5}}(2,1), \frac{1}{\sqrt{5}}(1,-2)\}$ is an orthonormal set and, since $\dim K^2 = 2$, an orthonormal basis of K^2.

(2) The canonical basis $\{e_1, \ldots, e_n\}$ of K^n is an orthonormal basis with respect to the standard inner product. It is easy to see that $\langle e_i, e_j \rangle = 0$ whenever $i \neq j$ and that

$$\|e_i\| = \sqrt{0^2 + \cdots + 0^2 + 1^2 + 0^2 + \cdots + 0^2} = 1 \qquad \forall i \in \{1, 2, \ldots, n\}.$$

(3) Consider $(C^0([0, 2\pi], \mathbb{R}), \langle \cdot, \cdot \rangle)$ with the inner product from Example 10.3.(3). A lengthy computation of integrals shows that $\{f_0\} \cup \{f_k \mid k \in \mathbb{N}\} \cup \{g_k \mid k \in \mathbb{N}\}$ is an orthonormal family in $C^0([0, 2\pi], \mathbb{R})$, where

$$f_0(x) = \frac{1}{2\pi}, \quad f_k(x) = \frac{1}{\sqrt{\pi}} \cos(kx), \quad g_k(x) = \frac{1}{\sqrt{\pi}} \sin(kx) \quad \forall k \in \mathbb{N},\ x \in [0, 2\pi].$$

The following result establishes the fact that the vectors in an orthogonal family indeed "point in different directions".

Theorem 10.20. *Let $(V, \langle \cdot, \cdot \rangle)$ be a Euclidean or a unitary vector space. Every orthogonal family $\{v_1, \ldots, v_k\}$ in V with $v_i \neq 0$ for each $i \in \{1, 2, \ldots, k\}$ is linearly independent.*

Proof. We show only the unitary case. Let $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$, such that $\sum_{i=1}^k \lambda_i v_i = 0$. Using the bilinearity properties of the inner product, we then compute

$$0 = \left\langle \sum_{i=1}^k \lambda_i v_i, \sum_{j=1}^k \lambda_j v_j \right\rangle = \sum_{i=1}^k \lambda_i \left\langle v_i, \sum_{j=1}^k \lambda_j v_j \right\rangle = \sum_{i=1}^k \sum_{j=1}^k \lambda_i \overline{\lambda}_j \langle v_i, v_j \rangle .$$

Since the family is orthogonal, the only non-vanishing summands in the above sum are those with $i = j$. Thus, it follows that

$$0 = \sum_{i=1}^k \lambda_i \overline{\lambda}_i \langle v_i, v_i \rangle = \sum_{i=1}^k |\lambda_i|^2 \langle v_i, v_i \rangle .$$

Since $\langle v_i, v_i \rangle > 0$ for all $i \in \{1, 2, \ldots, k\}$ this implies that $|\lambda_i|^2 = 0$ and thus $\lambda_i = 0$ for every i. Hence, $\{v_1, \ldots, v_k\}$ is linearly independent. □

The canonical basis of K^n has the advantage the coefficients of a vector given as a linear combination of $\{e_1, \ldots, e_n\}$ are given by the coordinate entries. In other words, these coefficients can be immediately read off. This property generalizes to orthonormal bases in the form that there is a simple way to compute the coefficients of a vector as a linear combination of orthonomal basis vectors.

Theorem 10.21. *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean or unitary vector space. Let $\{b_1, \dots, b_n\}$ be an orthonormal basis of V. Then*

$$v = \sum_{i=1}^{n} \langle v, b_i \rangle \, b_i \qquad \forall v \in V.$$

Proof. Let $K = \mathbb{R}$ if V is Euclidean and $K = \mathbb{C}$ if V is unitary. Let $v \in V$ and let $\lambda_1, \dots, \lambda_n \in K$, such that $v = \sum_{i=1}^{n} \lambda_i b_i$. Then for each $j \in \{1, 2, \dots, n\}$:

$$\langle v, b_j \rangle = \sum_{i=1}^{n} \left\langle \sum_{i=1}^{n} \lambda_i b_i, b_j \right\rangle = \sum_{i=1}^{n} \lambda_i \langle b_i, b_j \rangle = \lambda_j,$$

since $\langle b_i, b_j \rangle = \delta_{ij}$ as the basis is orthonormal. This shows the claim. □

While this is a useful property of orthonormal bases, we so far don't know if a given Euclidean or unitary vector space even possesses one. We will see that not only every such vector space admits an orthonormal basis, but that there is a constructive method to transform any given basis of a Euclidean or unitary vector space into an orthonormal basis. Before we investigate this result, we discuss the following technical lemma that will be required in its proof.

Lemma 10.22. *Let V be a Euclidean or a unitary vector space. Let $\{u_1, \dots, u_k\}$ be an orthonormal family and put $U := \operatorname{span}(\{u_1, \dots, u_k\})$. Then for all $v \in V$ it holds that*

$$\left(v - \sum_{i=1}^{k} \langle v, u_i \rangle \, u_i \right) \perp U.$$

Proof. For each $j \in \{1, 2, \dots, k\}$ we compute from the orthonormality of $\{u_1, \dots, u_k\}$ and the properties of inner products that

$$\left\langle v - \sum_{i=1}^{k} \langle v, u_i \rangle \, u_i, u_j \right\rangle = \langle v, u_j \rangle - \sum_{i=1}^{k} \langle v, u_i \rangle \langle u_i, u_j \rangle = \langle v, u_j \rangle - \langle v, u_j \rangle = 0,$$

hence $(v - \sum_{i=1}^{k} \langle v, u_i \rangle \, u_i) \perp u_j$ for every $j \in \{1, 2, \dots, k\}$. Since $U = \operatorname{span}(\{u_1, \dots, u_k\})$, this shows the claim. □

Theorem 10.23 (Gram-Schmidt orthonormalization process)**.** *Let V be a Euclidean or unitary vector space and let $\{v_1, \dots, v_k\}$ be linearly independent. Then $\{w_1, \dots, w_k\}$ is an orthonormal family with $\operatorname{span}(\{v_1, \dots, v_i\}) = \operatorname{span}(\{w_1, \dots, w_i\})$ for each $i \in \{1, 2, \dots, k\}$, where*

$$w_1 := \frac{v_1}{\|v_1\|}, \qquad w_{i+1} = \frac{v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle \, w_j}{\|v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle \, w_j\|} \qquad \forall i \in \{1, 2, \dots, k-1\}.$$

In particular, if V is finite-dimensional and $\{v_1, \dots, v_k\}$ is a basis of V, then $\{w_1, \dots, w_k\}$ is an orthonormal basis of V.

Proof. We want to prove this by induction. For the base case, we observe that

$$\langle w_1, w_1 \rangle = \left\langle \frac{v_1}{\|v_1\|}, \frac{v_1}{\|v_1\|} \right\rangle = \frac{\langle v_1, v_1 \rangle}{\|v_1\|^2} = 1,$$

hence $\{w_1\}$ is a (very small) orthonormal family. Moreover, since w_1 is a scalar multiple of v_1, it holds that $\operatorname{span}(\{v_1\}) = \operatorname{span}(\{w_1\})$, which completes the base hase.
Assume as induction hypothesis that $\{w_1, \dots, w_i\}$ is an orthonormal family for some $i < k$ with $\operatorname{span}(\{v_1, \dots, v_i\}) = \operatorname{span}(\{w_1, \dots, w_i\})$ and let w_{i+1} be defined as in the statement of the theorem. By Lemma 10.22, it holds that $w_{i+1} \perp w_j$ for each $j \in \{1, 2, \dots, i\}$. Moreover,

$$\|w_i\| = \left\| \frac{v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j}{\|v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j\|} \right\| = \frac{\|v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j\|}{\|v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j\|} = 1.$$

Adding these observations to the induction hypothesis shows that $\{w_1, \dots, w_i, w_{i+1}\}$ is an orthonormal family. By definition of w_{i+1}, it further holds that w_{i+1} is a linear combination of $\{w_1, \dots, w_i, v_{i+1}\}$. By the induction hypothesis, this yields

$$w_{i+1} \in \operatorname{span}(\{w_1, \dots, w_i, v_{i+1}\}) = \operatorname{span}(\{v_1, \dots, v_i v_{i+1}\}).$$

Hence, $\operatorname{span}(\{w_1, \dots, w_{i+1}\}) \subset \operatorname{span}(\{v_1, \dots, v_i, v_{i+1}\})$. To show the opposite inequality, we observe from the definition of w_{i+1} that, with $c := \|v_{i+1} - \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j\| \in \mathbb{R}$,

$$v_{i+1} = c w_{i+1} + \sum_{j=1}^{i} \langle v_{i+1}, w_j \rangle w_j \in \operatorname{span}(\{w_1, \dots, w_i, w_{i+1}\}),$$

hence by the induction hypothesis $\operatorname{span}(\{v_1, \dots, v_i, v_{i+1}\}) \subset \operatorname{span}(\{w_1, \dots, w_i, w_{i+1}\})$. This completes the induction step and thereby shows the claim. □

An example for the Gram-Schmidt process will be discussed in the exercise class on June 4th.

Corollary 10.24. *Every finite-dimensional Euclidean or unitary vector space possesses an orthonormal basis.*

Proof. Given an arbitrary basis, the process from Theorem 10.23 produces an orthonormal basis. □

We next want to see that orthogonality can be used to find complements of linear subspaces.

Definition 10.25. Let V be a Euclidean or a unitary vector space and let $M \subset V$ be a subset. We call

$$M^\perp := \{v \in V \mid \langle v, u \rangle = 0 \; \forall u \in M\}$$

the *orthogonal complement of* M. We further write $v^\perp := \{v\}^\perp$ for $v \in V$.

Theorem 10.26. *Let V be a Euclidean or unitary vector space.*

a) Let $M \subset V$. Then $M^\perp$ is a linear subspace of V.

b) If $U \subset V$ is a finite-dimensional linear subspace, then $V = U \oplus U^\perp$.

Proof. a) It obviously holds that $0 \in M^\perp$, so $M^\perp \neq \varnothing$. Let $v_1, v_2 \in M^\perp$. By the bilinearity of inner products, it holds for all $u \in M$ that

$$\langle v_1 + v_2, u \rangle = \langle v_1, u \rangle + \langle v_2, u \rangle = 0,$$

hence $v_1, v_2 \in M^\perp$. We further compute for all $\lambda \in K$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$) and $v \in M^\perp$ that

$$\langle \lambda v, u \rangle = \lambda \langle v, u \rangle = 0 \qquad \forall u \in M,$$

hence $\lambda v \in M^\perp$. This shows that $M^\perp$ is a linear subspace of M.

b) If $v \in U \cap U^\perp$, then by definition of $U^\perp$, it follows in particular that $\langle v, v\rangle = 0$. By the positive definiteness of inner products, this yields $v = 0$, so $U \cap U^\perp = \{0\}$. Let $k := \dim U$ and let $B = \{b_1, \dots, b_k\}$ be an orthonormal basis of U, which exists by Corollary 10.24. For $v \in V$ put $v_U := \sum_{i=1}^k \langle v, b_i\rangle\, b_i$. By Lemma 10.22, $v - v_U \in U^\perp$, hence $v = v_U + (v - v_U) \in U + U^\perp$. Thus, we have shown that $V = U \oplus U^\perp$.

□

Example 10.27. (1) For every Euclidean or unitary vector space, it holds that $V^\perp = \{0\}$ and $\{0\}^\perp = V$.

(2) Let $(2,1,1) \in \mathbb{R}^3$ and let $U := \mathbb{R}(2,1,1)$. One checks that $(2,1,1) \perp \{(1,-2,0),(1,0,-2)\}$ with respect to the standard inner product, hence $(1,-2,0),(1,0,-2) \in U^\perp$. By Theorem 10.26, $\dim U^\perp = 3 - \dim U = 2$, hence $U^\perp = \operatorname{span}(\{(1,-2,0)(1,0,-2)\})$ as the two vectors are evidently linearly independent.

We want to conclude this section by discussing a famous result by Riesz, which says that in the finite-dimensional case, an inner product can be used to construct an isomorphism between a real vector space and its dual space.

Theorem 10.28. *Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean vector space.*

a) For $w \in V$ let $\varphi_w : V \to K$, $\varphi(v) = \langle v, w\rangle$. Then $\varphi_w \in V^*$ and $\ker \varphi_w = w^\perp$ for each $w \in V$.

b) The map $F : V \to V^*$, $F(w) = \varphi_w$, is linear and injective.

c) (Riesz representation theorem) *If V is finite-dimensional, then $F : V \to V^*$ is an isomorphism.*

Proof. a) This follows immediately from property (i) of inner products and from the definition of orthogonal complements.

b) Let $w_1, w_2 \in V$. Then for each $v \in V$ we compute

$$\begin{aligned}(F(w_1 + w_2))(v) &= \varphi_{w_1+w_2}(v) = \langle v, w_1 + w_2\rangle \\ &= \langle v, w_1\rangle + \langle v, w_2\rangle = \varphi_{w_1}(v) + \varphi_{w_2}(v) = (F(w_1))(v) + (F(w_2))(v).\end{aligned}$$

Hence, $F(w_1 + w_2) = F(w_1) + F(w_2)$. For all $\lambda \in \mathbb{R}$ and $v \in V$ we further compute

$$F(\lambda w_1)(v) = \langle v, \lambda w_1\rangle = \lambda \langle v, w_1\rangle = \lambda(F(w_1))(v),$$

which yields $F(\lambda w_1) = \lambda F(w_1)$. Thus, F is linear.

Concerning injectivity, let $w_1, w_2 \in V$, such that $F(w_1) = F(w_2)$. Then

$$\langle v, w_1\rangle = \langle v, w_2\rangle \quad \Leftrightarrow \quad \langle v, w_1 - w_2\rangle = 0 \quad \forall v \in V.$$

Inserting $v = w_1 - w_2$ then implies $\langle w_1 - w_2, w_1 - w_2\rangle = 0$, so by positive definiteness of the inner product, $w_1 - w_2 = 0$, i.e. $w_1 = w_2$. Hence, F is injective.

c) If V is finite-dimensional, then $\dim V = \dim V^*$ by Theorem 7.28. Since F is linear and injective by b), it follows by Corollary 7.23.c) that F is an isomorphism.

□

Remark 10.29. (1) The Riesz representation theorem has a direct analogue for unitary vector spaces. More precisely, if $(V, \langle\cdot,\cdot\rangle)$ is a finite-dimensional unitary vector space, then for each $\varphi \in V^*$ there exists $w \in V$, such that $\varphi(v) = \langle v, w\rangle$ for each $v \in V$. Since in this case, the map F from Theorem 10.28 is not linear, but semilinear, this requires a little additional work and the reader is invited to carry out the details of that proof.

(2) Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean vector space and let $F : V \to V^*$ be the isomorphism from Theorem 10.28.c). There is an explicit formula for the vector representing a linear form in terms of a basis. If $\{b_1, \ldots, b_n\}$ is an orthonormal basis of V, then

$$F^{-1}(\varphi) = \sum_{i=1}^{n} \varphi(b_i) b_n.$$

An explicit computation shows that this vector really has the desired property.

(3) The statement of the Riesz representation theorem is in general false for infinite-dimensional vector spaces. Consider the real vector space $C^0([0,1], \mathbb{R})$ and the linear form

$$\delta \in (C^0([0,1], \mathbb{R}))^*, \quad \delta(f) = f(0).$$

Assume there was $g \in C^0([0,1], \mathbb{R})$, such that

$$\delta(f) = \langle f, g \rangle \quad \Leftrightarrow f(0) = \int_0^1 f(x) g(x)\, dx \quad \forall f \in C^0([0,1], \mathbb{R}).$$

Inserting $f(x) := xg(x)$ would then yield

$$\int_0^1 x(g(x))^2\, dx = 0,$$

which by Exercise 1 from Sheet 11 from Mathematics 1 would imply $x(g(x))^2 = 0$ and thus $g(x) = 0$ for all $x \in [0,1]$. But then $\delta(f) = 0$ for all $f \in C^0([0,1], \mathbb{R})$, which is a contradiction. Hence, such a g does not exist.

10.3 Orthogonal maps and matrices

In this section, we are going to study linear maps between Euclidean or unitary that are "compatible" with respect to the inner products. This generalizes the notion of a linear map which preserves lengths of and angles between vectors in $\mathbb{R}^n$.

Note: In this section, we will write often put the unitary cases in parentheses. Whenever this is the case, this indicates that the sentence or the paragraph remain true if one replaces every term by the one in the parentheses immediately thereafter.

Definition 10.30. Let $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot \rangle_W)$ be two Euclidean (or two unitary) vector spaces.

a) A linear map $f : V \to W$ is called *orthogonal* (*unitary*) if

$$\langle f(v), f(w) \rangle_W = \langle v, w \rangle_V \quad \forall v, w \in V.$$

b) If V is Euclidean (unitary), then the set of all orthogonal (unitary) endomorphisms of V is denoted by $O(V)$ (by $U(V)$).

c) An isomorphism $f : V \to W$ which is also orthogonal (unitary) is called an *isometric isomorphism* or simply an *isometry*.

d) Two Euclidean (or two unitary) vector spaces V and W are called *isometric* if there exists an isometry $f : V \to W$.

Remark 10.31. Let $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot, \rangle_W)$ be two Euclidean (two unitary) vector spaces.

(1) Any orthogonal (unitary) map $f : V \to W$ is injective: if $v \in \ker f$, then

$$\langle v, v \rangle_V = \langle f(v), f(v)(>_W = \langle 0, 0 \rangle_W = 0,$$

hence $v = 0$ by the positive definiteness of inner products. Hence, $\ker f = \{0\}$, so f is injective by Theorem 7.10.b).

(2) If V is finite-dimensional, then any orthogonal (unitary) endomorphism $f : V \to V$ is an isometry. This follows from (1) and Corollary 7.23.c).

(3) Assume that V is Euclidean. Any orthogonal map $f : V \to V$ is *length-preserving* and *angle-preserving* in the following sense:

If $f \in L(V, V)$ is orthogonal (unitary), then $\|f(v)\| = \sqrt{\langle f(v), f(v) \rangle} = \sqrt{\langle v, v \rangle} = \|v\|$ for all $v \in V$. It further follows that

$$\angle(f(v), f(w)) = \angle(v, w) \qquad \forall v, w \in V \setminus \{0\}.$$

The following result collects some basic facts about orthogonal and unitary maps.

Theorem 10.32. *Let $(V, \langle \cdot, \cdot \rangle_V)$, $(W, \langle \cdot, \cdot \rangle_W)$ and $(X, \langle \cdot, \cdot \rangle_X)$ be three Euclidean (three unitary) vector spaces.*

a) If $f : V \to W$ and $g : W \to X$ are orthogonal (unitary), then $g \circ f : V \to X$ is orthogonal (unitary).

b) If $f : V \to V$ is orthogonal (unitary), then $f^{-1} : V \to V$ is unitary.

c) If $f : V \to V$ is orthogonal (unitary) and if $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$) is an eigenvalue of f, then $|\lambda| = 1$.

Proof. (This proof is not discussed in the lecture notes.)

a) This is obvious.

b) We compute for all $v, w \in V$ that

$$\langle v, w \rangle_V = \left\langle f(f^{-1}(v)), f(f^{-1}(w)) \right\rangle_V \overset{f \text{ orth.}}{=} \left\langle f^{-1}(v), f^{-1}(w) \right\rangle_V .$$

Hence, f^{-1} is orthogonal (unitary).

c) If $\lambda \in K$ is an eigenvalue of f and if $v \in V$ is an eigenvector of f corresponding to λ, then $\|f(v)\| = \|v\|$ by Remark 10.31.(2) and $\|f(v)\| = \|\lambda v\| = |\lambda| \|$ by the properties of a norm, hence $\|v\| = |\lambda| \|v\|$. Since $v \neq 0$, this yields $|\lambda| = 1$.

□

We next want to investigate the behaviour of orthogonal maps on orthonormal bases.

Theorem 10.33. *Let $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot \rangle_W)$ be two Euclidean (or two unitary) vector spaces. Assume that V is finite-dimensional and let $\{b_1, \ldots, b_n\}$ be an orthonomal basis of V. Let $f : V \to W$ be linear. Then f is orthogonal (unitary) if and only if $\{f(b_1), \ldots, f(b_n)\}$ with $f(b_i) \neq f(b_j)$ for $i \neq j$ is an orthonormal family in W.*

Proof. "⇒": If f is orthogonal (unitary), then

$$\langle f(b_i), f(b_j)\rangle_W = \langle b_i, b_j\rangle_V = \delta_{ij} \quad \forall i,j \in \{1,2,\dots,n\},$$

hence $\{f(b_1),\dots,f(b_n)\}$ is an orthonormal family in W.
"⇐": If $\{f(b_1),\dots,f(b_n)\}$ is an orthonormal family in W with $f(b_i) \neq f(b_j)$ for $i \neq j$, then we obtain for all $v = \sum_{i=1}^n \lambda_i b_i$, $w = \sum_{j=1}^n \mu_j b_j \in V$ with the properties of inner products that

$$\begin{aligned}\langle f(v), f(w)\rangle_W &= \left\langle \sum_{i=1}^n \lambda_i f(b_i), \sum_{j=1}^n \mu_j f(b_j)\right\rangle_W = \sum_{i=1}^n\sum_{j=1}^n \lambda_i \overline{\mu}_j \langle f(b_i), f(b_j)\rangle_W \\ &= \sum_{i=1}^n\sum_{j=1}^n \lambda_i \overline{\mu}_j \langle b_i, b_j\rangle_V = \left\langle \sum_{i=1}^n \lambda_i b_i, \sum_{j=1}^n \mu_j b_j\right\rangle_V = \langle v, w\rangle_V,\end{aligned}$$

which shows that f is orthogonal if V is Euclidean and unitary if V is unitary. □

The previous theorem has an interesting consequence. It shows that a finite-dimensional Euclidean or unitary vector space is not only isomorphic to K^n for suitable n, but even isometric.

Corollary 10.34. *Let $(V, \langle\cdot,\cdot\rangle_V)$ be a finite-dimensional Euclidean or unitary vector space and let $n = \dim V$. Then V is isometric to K^n with the standard inner product, where $K = \mathbb{R}$ if V is Euclidean and $K = \mathbb{C}$ if V is unitary.*

Proof. Let $B = (b_1,\dots,b_n)$ be an orthonormal basis of V which exists by Theorem 10.23 and let $\phi_B : V \to K^n$ be its coordinate representation. We have seen that ϕ_B is an isomorphism. Since $\phi_B(b_i) = e_i$, we obtain $\{\phi_B(b_1),\dots,\phi_B(b_n)\} = \{e_1,\dots,e_n\}$, which is an orthonormal family in K^n. Hence, ϕ_B is orthogonal if V is Euclidean and unitary if V is unitary by Theorem 10.33. □

Corollary 10.35. *If V and W are two Euclidean (or two unitary) vector spaces with $\dim V = \dim W$, then V and W are isometric.*

Proof. Let $K = \mathbb{R}$ if V and W are Euclidean and $K = \mathbb{C}$ if V and W are unitary. By Corollary 10.34 there are isometries $f : V \to K^n$ and $g : W \to K^n$. Then, by Theorem 10.32, $g^{-1} \circ f : V \to W$ is an isometry, so V and W are isometric. □

The notion of orthogonal and unitary maps transfers to real and complex matrices as well.

Definition 10.36. Let $n \in \mathbb{N}$.

(1) Let $\mathbb{R}^n$ be equipped with the standard inner product. A matrix $A \in M(n \times n, \mathbb{R})$ is called *orthogonal* if

$$\langle Av, Aw\rangle = \langle v, w\rangle \quad \forall v, w \in \mathbb{R}^n.$$

The set of all orthogonal $n \times n$-matrices is denoted by $O(n)$ and called *the orthogonal group in dimension n*.

(2) Let $\mathbb{C}^n$ be equipped with the standard inner product. A matrix $A \in M(n \times n, \mathbb{C})$ is called *unitary* if

$$\langle Av, Aw\rangle = \langle v, w\rangle \quad \forall v, w \in \mathbb{C}^n.$$

The set of all unitary $n \times n$-matrices is denoted by $U(n)$ and called *the unitary group in dimension n*.

Remark 10.37. (1) Note that by definition $A \in M(n \times n, K)$ is orthogonal (unitary) if and only if $f_A : K^n \to K^n$ is orthogonal (unitary) with respect to the standard inner product.

(2) In analogy with Remark 10.31.a), it follows that

$$O(n) \subset \mathrm{GL}(n, \mathbb{R}) \qquad \text{and} \qquad U(n) \subset \mathrm{GL}(n, \mathbb{C}) \qquad \text{for all } n \in \mathbb{N}.$$

One further derives from Theorem 10.32 applied to maps of the form f_A, where $A \in O(n)$, that $O(n)$ and $U(n)$ are indeed groups with respect to the matrix product.

Next we establish a not too surprising relation between the orthogonality of a linear maps and the orthogonality of its matrix representations.

Theorem 10.38. *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean (unitary) vector space, B be an orthonormal basis of V and $f : V \to V$ be an endomorphism. Then f is orthogonal (unitary) if and only if $M_B(f)$ is orthogonal (unitary).*

Proof. "⇒": Let $n := \dim V$ and $A := M_B(f)$. As seen in the proof of Corollary 10.34, since B is an orthonormal basis, the coordinate representation $\Phi_B : V \to K^n$ is an isometry, hence $f_A = \Phi_B \circ f \circ \Phi_B^{-1}$ is orthogonal (unitary) by Theorem 10.32.
"⇐": Analogously, if A and thus f_A is orthogonal (unitary), then $f = \Phi_B^{-1} \circ f_A \circ \Phi_B$ is orthogonal (unitary) by Theorem 10.32. □

Comparing the product of matrices with the standard inner product on K^n, one realizes that matrix products can be fully described in terms of this inner product in the following way:

Let $\ell, m, n \in \mathbb{N}$ and let $A \in M(\ell \times m, K)$ and $B \in M(m \times n, K)$. If $(r_1, \ldots, r_\ell)$ are the row vectors of A, seen as elements of K^m, and $(c_1, \ldots, c_n)$ are the column vectors of B, also seen as elements of K^m, then by definition of the matrix product, one checks that $AB \in M(\ell \times n, K)$ is given by

$$AB = (\langle r_i, c_j \rangle) = \begin{pmatrix} \langle r_1, c_1 \rangle & \langle r_1, c_2 \rangle & \ldots & \langle r_1, c_n \rangle \\ \langle r_2, c_1 \rangle & \langle r_2, c_2 \rangle & \ldots & \langle r_2, c_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle r_\ell, c_1 \rangle & \langle r_\ell, c_2 \rangle & \ldots & \langle r_\ell, c_n \rangle \end{pmatrix}. \tag{10.2}$$

This observation is not only a helpful way of memorizing the definition of the matrix product, but also immensely helpful to establish properties of orthogonal and unitary matrices in the following theorems.

Theorem 10.39. *Let $n \in \mathbb{N}$ and let $A \in M(n \times n, \mathbb{R})$. The following statements are equivalent:*

(i) A is orthogonal.

(ii) The columns of A form an orthonormal basis of $\mathbb{R}^n$.

(iii) $A^t A = I_n$.

(iv) A is invertible and $A^{-1} = A^t$.

(v) The rows of A form an orthonormal basis of $\mathbb{R}^n$.

Proof. (i) $\Rightarrow$ (ii): We have seen that the columns of A are given by $(Ae_1, \ldots, Ae_n)$. Since $(e_1, \ldots, e_n)$ is an orthonormal basis and A is orthogonal, it follows from Theorem 10.33 that the columns of A form an orthonormal basis of $\mathbb{R}^n$.
(ii) $\Rightarrow$ (i): By assumption, $\langle Ae_i, Ae_j \rangle = \delta_{ij}$ for all $i, j \in \{1, 2, \ldots, n\}$. Let $v = (\lambda_1, \ldots, \lambda_n), w = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$. Then

$$\begin{aligned}
\langle Av, Aw \rangle &= \left\langle A\left(\sum_{i=1}^{n} \lambda_i e_i\right), A\left(\sum_{j=1}^{n} \mu_j e_j\right) \right\rangle \\
&= \left\langle \sum_{i=1}^{n} \lambda_i Ae_i, \sum_{j=1}^{n} \mu_j Ae_j \right\rangle \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j \langle Ae_i, Ae_j \rangle \\
&= \sum_{i=1}^{n} \lambda_i \mu_i = \langle (\lambda_1, \ldots, \lambda_n), (\mu_1, \ldots, \mu_n) \rangle = \langle v, w \rangle .
\end{aligned}$$

Hence, A is orthogonal.
(ii) $\Leftrightarrow$ (iii): Let $(c_1, \ldots, c_n)$ be the column vectors of A. Then (ii) is equivalent to $\langle c_i, c_j \rangle = \delta_{ij}$ for all $i, j \in \{1, 2, \ldots, n\}$. Since the row vectors of A^t are given by $(c_1, \ldots, c_n)$ (up to identifying row vectors with column vectors), we derive from (10.2) that this is in turn equivalent to $A^t A = (\langle c_i, c_j \rangle) = (\delta_{ij}) = I_n$.
(iii) $\Rightarrow$ (iv): Since $f_{A^t} \circ f_A = \mathrm{id}_{\mathbb{R}^n}$, the map f_A is linear and injective, hence an isomorphism by Corollary 7.23.c). Thus, A is invertible. Using this observation, we further derive from $A^t A = I_n$ that

$$AA^t A = A \overset{\text{Theorem 9.13}}{\Leftrightarrow} (AA^t - I_n)A = 0 \overset{A \text{ invertible}}{\Leftrightarrow} AA^t - I_n = 0 \quad \Leftrightarrow \quad AA^t = I_n.$$

Hence, A is invertible with $A^{-1} = A^t$.
(iv) $\Rightarrow$ (iii): This is obvious.
(iv) $\Rightarrow$ (v): This follows in complete analogy with "(iii) $\Rightarrow$ (ii)", using that the column vectors of A^t are the row vectors of A.
(v) $\Rightarrow$ (iv): In analogy with "(ii) $\Rightarrow$ (iii)" it follows that $AA^t = I_n$. In analogy with "(iii) $\Rightarrow$ (iv)", but with the roles of A and A^t reversed, one derives that $A^t A = I_n$ as well, which proves (iv). □

There is an analogue of Theorem 10.39 for unitary matrices. Since its proof is given along the lines of the proof of Theorem 10.39, we omit its proof.
Given $A = (a_{ij}) \in M(n \times n, \mathbb{C})$ we define $\overline{A} := (\overline{a}_{ij}) \in M(n \times n, \mathbb{C})$, i.e. the (i, j)-entry of A is the complex conjugate of a_{ij} for all $i, j \in \{1, 2, \ldots, n\}$.

Theorem 10.40. *Let $n \in \mathbb{N}$ and let $A \in M(n \times n, \mathbb{C})$. The following statements are equivalent:*

(i) A is orthogonal.

(ii) The columns of A form an orthonormal basis of $\mathbb{C}^n$.

(iii) $\overline{A}^t A = I_n$.

(iv) A is invertible and $A^{-1} = \overline{A}^t$.

(v) The rows of A form an orthonormal basis of $\mathbb{C}^n$.

The previous theorems show that we can make strong general statements about determinant of orthogonal and unitary maps.

Corollary 10.41. *Let $n \in \mathbb{N}$.*

a) If $A \in O(n)$, then $\det A \in \{-1, 1\}$.

b) If $A \in U(n)$, then $|\det A| = 1$.

Proof. a) We compute that

$$\det A \cdot \det A^t \overset{\text{Theorem 9.14}}{=} \det(AA^t) \overset{\text{Theorem 10.39}}{=} \det I_n = 1.$$

By Theorem 9.17, this yields $(\det A)^2 = 1$ and thus $|\det A| = 1$. Since $\det A \in \mathbb{R}$, this yields $\det A \in \{-1, 1\}$.

b) From the explicit form of the determinant in (9.1), the properties of complex conjugation and Theorem 9.17, one observes that $\det \overline{A}^t = \overline{\det A^t} = \overline{\det A}$. Thus, by Theorems 9.17 and 10.40,

$$1 = \det I_n = \det(A\overline{A}^t) = \det A \cdot \det \overline{A}^t = \det A \cdot \overline{\det A} = |\det A|^2.$$

Hence, $|\det A| = 1$.

□

The set of orthogonal endomorphisms of the Euclidean plane $\mathbb{R}^2$ with respect to the standard inner product can be determined explicitly. We will do so by equivalently determining the orthogonal group in dimension two and afterwards give a geometric interpretation of the occuring matrices.

Theorem 10.42. *The set of orthogonal 2×2-matrices is given by*

$$O(2) = \left\{ R_\alpha := \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \,\middle|\, \alpha \in [0, 2\pi) \right\} \cup \left\{ S_\alpha := \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix} \,\middle|\, \alpha \in [0, 2\pi) \right\}.$$

Proof. By Theorem 10.39, it holds that $O(2) = \{A \in M(2 \times 2, \mathbb{R}) \,|A^t A = I_2\}$.
An explicit computation of matrix products then shows that $R_\alpha, S_\alpha \in O(2)$ for each $\alpha \in [0, 2\pi)$, so it remains to show that every element of $O(2)$ is of the form R_α or of the form S_α.
Given $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in O(2)$, the condition $A^t A = I_2$ becomes

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

Since all real solutions of $x^2 + y^2 = 1$ are of the form $x = \cos\alpha$ and $y = \sin\alpha$ for some $\alpha \in [0, 2\pi)$, we derive from the first two equations that there are $\alpha, \beta \in [0, 2\pi)$, such that

$$a = \cos\alpha, \quad c = \sin\alpha, \quad b = \sin\beta, \quad d = \cos\beta.$$

The third equation then reads as

$$0 = \cos\alpha \sin\beta + \sin\alpha \cos\beta = 2\sin(\alpha + \beta),$$

where we used Theorem 3.78. Thus, by Corollary 3.87.a) there exists $k \in \mathbb{Z}$, such that $\alpha + \beta = k\pi$. Since $\alpha, \beta \in [0, 2\pi)$ it holds that $k \in \{0, 1, 2, 3\}$. We have to distinguish between two cases:

- If $k \in \{1,3\}$, i.e. if $\beta \in \{\pi - \alpha, 3\pi - \alpha\}$, then by the periodicity properties of sine and cosine
$$\cos\beta = \cos(\pi - \alpha) = -\cos\alpha, \quad \sin\beta = \sin(\pi - \alpha) = \sin\alpha.$$
Thus, in this case $A = S_\alpha$.

- If $k \in \{0,2\}$, i.e. if $\beta = \alpha = 0$ or $\beta = 2\pi - \alpha$, then
$$\cos\beta = -\cos(2\pi - \alpha) = \cos\alpha, \quad \sin\beta = \sin(2\pi - \alpha) = -\sin\alpha.$$
Thus, in this case $A = R_\alpha$.

This shows that every $A \in O(2)$ is of the desired form. □

Remark 10.43. The matrices R_α and S_α from Theorem 10.42 have a geometric meaning, but we unfortunately do not have enough time in the course to discuss it in detail.

(i) The map $f_{R_\alpha} : \mathbb{R}^2 \to \mathbb{R}^2$ induced by R_α, where $\alpha \in [0, 2\pi)$, is a counterclockwise rotation around the origin by the angle α. One observes (this is left to the reader as an exercise) that R_α does not any eigenvalues unless $\alpha \in \{0, \pi\}$.

(ii) The map $f_{S_\alpha} : \mathbb{R}^2 \to \mathbb{R}^2$ induced by S_α is a reflection along a line. One checks that this line is given by $\mathbb{R}(\cos(\frac{\alpha}{2}), \sin(\frac{\alpha}{2}))$.[3]

Theorem 10.42 thus shows that *every orthogonal endomorphism of* $\mathbb{R}^2$ *is either a rotation around the origin or a reflection along a line.*

Concluding this section, we want to give a definition that is not too important for us for the rest of this course, but that you might encounter in various situations in physics.

Definition 10.44. Let $n \in \mathbb{N}$.

a) The *special orthogonal group of dimension n* is given by
$$SO(n) = \{A \in O(n) \mid \det A = 1\} = \{A \in M(n \times n, \mathbb{R}) \mid A^t A = I_n, \det A = 1\}.$$

b) The *special unitary group of dimension n* is given by
$$SU(n) = \{A \in U(n) \mid \det A = 1\} = \{A \in M(n \times n, \mathbb{C}) \mid \overline{A}^t A = I_n, \det A = 1\}.$$

Remark 10.45. (1) Using the observations that $O(n)$ and $U(n)$ are groups with respect to the matrix product and Theorem 9.14, one shows that $SO(n)$ and $SU(n)$ are again groups with respect to the matrix product.

(2) Note that by definition, $SO(n) = O(n) \cap \mathrm{SL}(n, \mathbb{R})$ and $SU(n) = U(n) \cap \mathrm{SL}(n, \mathbb{C})$, with respect to the special linear group defined in Corollary/Definition 9.16.

Example 10.46. With R_α and S_α given as in Theorem 10.42, one observes that $\det R_\alpha = \cos^2\alpha + \sin^2\alpha = 1$ and $\det S_\alpha = -\cos^2\alpha - \sin^2\alpha = -1$ for all $\alpha \in [0, 2\pi)$. Thus, by Theorem 10.42, it holds that
$$SO(2) = \left\{ \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \;\middle|\; \alpha \in [0, 2\pi) \right\}.$$

[3] An exercise that helps in understanding the geometry: Show that S_α is diagonalizable and compute an eigenbasis of $\mathbb{R}^2$ associated with S_α.

10.4 Self-adjoint endomorphisms

In our studies of finite-dimensional vector spaces we have so far seen two kinds of bases with useful properties: eigenbases and orthonormal bases. While an eigenbasis allows for a simple description of an endomorphism, an orthonormal basis has several useful properties and provides us with intuition about lengths and angles between vectors.
The optimal case would be to have all of these properties combined, i.e. to have an *orthonormal eigenbasis* with respect to a given endomorphism. In this section, we will introduce a class of endomorphisms which has precisely this property, namely self-adjoint endomorphisms. These maps occur naturally in several situations in physics, for example in the infinite-dimensional case (which we shall not discuss here) many observables in quantum mechanics like position, momentum and spin are described by self-adjoint operators.

Definition 10.47. Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean or a unitary vector space. An endomorphism $f : V \to V$ is *self-adjoint* if

$$\langle f(v), w\rangle = \langle v, f(w)\rangle \quad \forall v, w \in V.$$

Example 10.48. (1) Obviously, $\mathrm{id}_V : V \to V$ and the zero map are self-adjoint for any Euclidean or unitary vector space.

(2) $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = (2x, 3y)$ is self-adjoint, since

$$\begin{aligned}\langle f(x_1,y_1),(x_2,y_2)\rangle &= \langle (2x_1,3y_1),(x_2,y_2)\rangle = 2x_1x_2 + 3y_1y_2 = \langle (x_1,y_1),(2x_2,3y_2)\rangle \\ &= \langle (x_1,y_1), f(x_2,y_2)\rangle\end{aligned}$$

for all $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^2$.

(3) $g : \mathbb{R}^2 \to \mathbb{R}^2$, $g(x,y) = (y,-x)$ is not self-adjoint, since e.g.

$$\langle g(1,0),(0,1)\rangle = \langle (0,-1),(0,1)\rangle = -1 \neq 1 = \langle (1,0),(1,0)\rangle = \langle (1,0), g(0,1)\rangle .$$

(4) Let $h \in C^0([0,1],\mathbb{R})$ and consider $T : C^0([0,1],\mathbb{R}) \to C^0([0,1],\mathbb{R})$, $T(f) = h \cdot f$. One checks without difficulties that T is linear and we compute for the previously studied inner product for all $f, g \in C^0([0,1],\mathbb{R})$ that

$$\begin{aligned}\langle T(f), g\rangle = \langle h\cdot f, g\rangle &= \int_0^1 h(x)f(x)g(x)\,dx = \int_0^1 f(x)(h(x)g(x))\,dx \\ &= \langle f, h\cdot g\rangle = \langle f, T(g)\rangle .\end{aligned}$$

Hence, T is self-adjoint.

The following properties of self-adjoint endomorphisms will turn out to be more important than one might think upon first sight.

Theorem 10.49. *Let $(V, \langle\cdot,\cdot\rangle)$ be a Euclidean or a unitary vector space and let $f : V \to V$ be a self-adjoint endomorphism. Then:*

a) Every eigenvalue of f is a real number.

b) If $\lambda, \mu \in \mathbb{R}$ are eigenvalues of f with $\lambda \neq \mu$, then $E_\lambda(f) \perp E_\mu(f)$.

Proof. a) In the case that V is Euclidean, this is clear, so assume that f is unitary. Let $\lambda \in \mathbb{C}$ be an eigenvalue of f and let $v \in V \setminus \{0\}$ be an eigenvector to the eigenvalue λ. Since f is self-adjoint, we compute that

$$\lambda \langle v, v \rangle = \langle \lambda v, v \rangle = \langle f(v), v \rangle = \langle v, f(v) \rangle = \langle v, \lambda v \rangle = \overline{\lambda} \langle v, v \rangle .$$

Since $\langle v, v \rangle \neq 0$ by the positive definiteness of inner products, this yields $\lambda = \overline{\lambda}$ which holds if and only if $\lambda \in \mathbb{R}$.

b) Let $v \in E_\lambda(f)$ and $w \in E_\mu(f)$. Since f is self-adjoint, we compute that

$$\lambda \langle v, w \rangle = \langle \lambda v, w \rangle = \langle f(v), w \rangle = \langle v, f(w) \rangle = \langle v, \mu w \rangle = \mu \langle v, w \rangle$$

since $\mu \in \mathbb{R}$ by a). This yields $(\lambda - \mu) \langle v, w \rangle = 0$ and since $\lambda \neq \mu$, it follows that $\langle v, w \rangle = 0$. Since $v \in E_\lambda(f)$ and $w \in E_\mu(f)$ were chosen arbitrarily, this shows the claim.

□

Remark 10.50. While it is easy to see that sums and scalar multiples of self-adjoint endomorphisms by real numbers are again self-adjoint, the composition of two self-adjoint endomorphisms is *not* necessarily self-adjoint. if $f, g : V \to V$ are self-adjoint endomorphisms of a Euclidean or unitary vector space, then for all $v, w \in V$:

$$\langle (g \circ f)(v), w \rangle = \langle g(f(v)), w \rangle = \langle f(v), g(w) \rangle = \langle v, f(g(w)) \rangle = \langle v, (f \circ g)(w) \rangle .$$

So $g \circ f$ is self-adjoint if and only if $g \circ f = f \circ g$, which is not always the case.

We next want to look at matrix representations of self-adjoint endomorphisms. We introduce an explicit property for matrices and will show that it corresponds to self-adjointness when it comes to matrix representations.

Definition 10.51. Let $n \in \mathbb{N}$.

a) A quadratic real matrix $A = (a_{ij}) \in M(n \times n, \mathbb{R})$ is called *symmetric* if

$$A^t = A,$$

i.e. if $a_{ij} = a_{ji}$ for all $i, j \in \{1, 2, \ldots, n\}$.

b) A quadratic complex matrix $A = (a_{ij}) \in M(n \times n, \mathbb{C})$ is called *Hermitian*[4] if

$$\overline{A}^t = A,$$

i.e. if $a_{ij} = \overline{a}_{ji}$ for all $i, j \in \{1, 2, \ldots, n\}$.

Example 10.52. $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$ is symmetric and $\begin{pmatrix} 1 & 1+i & 2+i \\ 1-i & 2 & 3-2i \\ 2-i & 3+2i & 5 \end{pmatrix}$ is Hermitian.

Theorem 10.53. *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean or unitary vector space, $B = (b_1, \ldots, b_n)$ be an orthonormal basis of V and $f : V \to V$ be an endomorphism.*

a) In case V is Euclidean, f is self-adjoint if and only if $M_B(f)$ is symmetric.

[4]This word might look weird upon first sight. It it named after the French mathematician Charles Hermite (1822-1901), who gave the original proof of Theorem 10.49.a).

b) In case V is unitary, f is self-adjoint if and only if $M_B(f)$ is Hermitian.

Proof. We will only prove part b), since part a) is proven along the same lines. Put $A := M_B(f)$ and denote its entries by $A = (a_{ij})$.
"⇒": If f is self-adjoint then it particularly holds for all $i, j \in \{1, 2, \dots, n\}$ that

$$\langle f(b_i), b_j \rangle = \langle b_i, f(b_j) \rangle \overset{\text{Theorem 8.20}}{\Leftrightarrow} \left\langle \sum_{k=1}^{n} a_{ki} b_k, b_j \right\rangle = \left\langle b_i, \sum_{k=1}^{n} a_{kj} b_k \right\rangle$$

$$\Leftrightarrow \quad \sum_{k=1}^{n} a_{ki} \langle b_k, b_j \rangle = \sum_{k=1}^{n} \overline{a}_{kj} \langle b_i, b_k \rangle \quad \Leftrightarrow \quad \sum_{k=1}^{n} a_{ki} \delta_{kj} = \sum_{k=1}^{n} \overline{a}_{kj} \delta_{ik} \quad \Leftrightarrow \quad a_{ji} = \overline{a}_{ij}.$$

Thus, A is Hermitian.
"⇒": As shown in the opposite implication, if A is Hermitian, then

$$\langle f(b_i), b_j \rangle = \langle b_i, f(b_j) \rangle \quad \forall i, j \in \{1, 2, \dots, n\}. \tag{10.3}$$

Let $v = \sum_{i=1}^{n} \lambda_i b_i, w = \sum_{j=1}^{n} \lambda_j b_j \in V$. Then, by the linearity properties of inner products,

$$\begin{aligned} \langle f(v), w \rangle &= \left\langle \sum_{i=1}^{n} \lambda_i f(b_i), \sum_{j=1}^{n} \mu_j b_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \overline{\mu}_j \langle f(b_i), b_j \rangle \\ &\overset{(10.3)}{=} \sum_{i=1}^{n} \lambda_i \overline{\mu}_j \langle b_i, f(b_j) \rangle = \left\langle \sum_{i=1}^{n} \lambda_i b_i, \sum_{j=1}^{n} \mu_j f(b_j) \right\rangle = \langle v, f(w) \rangle . \end{aligned}$$

Hence, f is self-adjoint. □

As indicated in the introduction to this section, one motivation to study self-adjoint endomorphisms is the existence of orthonormal eigenbases for such maps. The following important theorem manifests this property.

Theorem 10.54 (spectral theorem)**.** *Let V be a finite-dimensional Euclidean or unitary vector space. Then for every self-adjoint endomorphism $f : V \to V$ there exists an orthonormal basis of V which consists of eigenvectors of f. In particular, f is diagonalizable.*

Proof. We first consider the case that V is unitary. We want to show the claim by induction over $n = \dim V$. The base case of $n = 1$ is obvious. Concerning the induction step, we assume as induction hypothesis that the claim holds for all self-adjoint endomorphisms of unitary vector spaces of dimension $n - 1$ for a certain $n \geq 2$.
Assume that $\dim V = n$ and let $f \in L(V, V)$ be self-adjoint. As a complex polynomial, χ_f decomposes into linear factors, so there are eigenvalues $\lambda_1, \dots, \lambda_n$, such that

$$\chi_f(t) = (-1)^n (t - \lambda_1)(t - \lambda_2) \dots (t - \lambda_n) \quad \forall t \in \mathbb{C}.$$

(By Theorem 10.49.a), $\lambda_i \in \mathbb{R}$ for all $i \in \{1, 2, \dots, n\}$.) Let $v_1 \in V$ be an eigenvector with $\|v_1\| = 1$ corresponding to λ_1 and put $W := v_1^{\perp} = (\mathbb{R} v_1)^{\perp}$. By Theorem 10.26 and Corollary 6.41, $\dim W = n - 1$. Moreover, for each $w \in W$ we compute that

$$\langle v_1, f(w) \rangle = \langle f(v_1), w \rangle = \langle \lambda_1 v_1, w \rangle = \lambda_1 \langle v_1, w \rangle = 0,$$

so $f(w) \in v_1^{\perp} = W$. Hence, f restricts to a self-adjoint endomorphism $f|_W : W \to W$. By the induction hypothesis, there exists an orthonormal basis $\{v_2, \dots, v_n\}$ of W consisting of eigenvectors of $f|_W$, hence of f. Since by construction $v_i \perp v_1$ for all $i \in \{2, 3, \dots, n\}$, the family $\{v_1, \dots, v_n\}$ is orthonormal and thus, since $\dim V = n$, an orthonormal basis of V.

In the case that V is Euclidean, we can apply the very same line of argument if we can show that χ_f decomposes into linear factors. Let B be an orthonormal basis of V and put $A := M_B(f) \in M(n \times n, \mathbb{R})$. Then $\chi_f = \chi_A$. By Theorem 10.53, A is a symmetric matrix. If we view its entries as complex numbers, i.e. view A as a matrix in $M(n \times n, \mathbb{C})$, then A is Hermitian as well. From the unitary case, we observe that as a complex polynomial χ_A decomposes into linear factors, such that each of the linear factors is a *real* polynomial. This shows that $\chi_A = \chi_f$ decomposes into linear factors as a real polynomial as well. Hence, one proceeds as in the unitary case to show the claim. □

Corollary 10.55 (spectral theorem for matrices)**.** *Let $n \in \mathbb{N}$.*

a) Let $A \in M(n \times n, \mathbb{R})$ be symmetric. Then there exists $S \in O(n)$, such that SAS^t is a diagonal matrix.

b) Let $A \in M(n \times n, \mathbb{C})$ be Hermitian. Then there exists $S \in U(n)$, such that $SA\overline{S}^t$ is a diagonal matrix.

Proof. Again, we shall only prove part b) as part a) follows along the same lines.
By the spectral theorem, $f_A : \mathbb{C}^n \to \mathbb{C}^n$ is diagonalizable and there exists an orthonormal basis $\{b_1, \ldots, b_n\}$ of $\mathbb{C}^n$ consisting of eigenvectors of A. By Theorem 9.36 there exists $S \in \mathrm{GL}(n, \mathbb{C})$, such that SAS^{-1} is diagonal. By construction of S in the proof of that theorem, this matrix can be chosen to satisfy $Sb_i = e_i$ for each $i \in \{1, 2, \ldots, n\}$, which yields $S \in U(n)$ by Theorem 10.33. Since $S^{-1} = \overline{S}^t$ by Theorem 10.40, it follows that $SA\overline{S}^t$ is diagonal for such a choice of S. □

Remark 10.56. There also exists a version of the spectral theorem for self-adjoint endomorphisms of *infinite-dimensional* vector spaces that is crucial to the study of Dirac operators and other operators from theoretical physics. However, the infinite-dimensional version requires a lot of prerequisites from functional analysis, so it is (yet) out of reach for us.

The following corollary summarizes the main results of this section in a concise way.

Corollary 10.57. *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean or unitary vector space.*
Let $f : V \to V$ be a self-adjoint endomorphism and let $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ be the pairwise distinct eigenvalues of f. Then

$$V = E_{\lambda_1}(f) \oplus E_{\lambda_2}(f) \oplus \cdots \oplus E_{\lambda_k}(f)$$

and any two eigenspaces to distinct eigenvalues are orthogonal to each other.

Proof. The direct sum decomposition follows directly from combining Theorem 10.54 with Theorem 9.41. The orthogonality of the eigenspaces was shown in Theorem 10.49.b). □

While Corollary 10.57 is rather theoretical, the proofs results from this chapter also deliver the "theoretical guarantee" that the following method can be used to explicitly find orthonormal eigenbases for self-adjoint endomorphisms $f : K^n \to K^n$ for $K = \mathbb{R}$ or $K = \mathbb{C}$.

Method for finding an orthonormal eigenbasis for a self-adjoint endomorphism $f : K^n \to K^n$

- Compute χ_f, find its decomposition into linear factors (which exists as seen above) and read off its eigenvalues.
- If $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ are the pairwise distinct eigenvalues of f, find a basis $\{b_1^i, \ldots, b_{n_i}^i\}$ of $E_{\lambda_i}(f)$ by solving the linear system $f(v) = \lambda_i v$.
- Apply the Gram-Schmidt orthonormalization process individually to each $\{b_1^i, \ldots, b_{n_i}^i\}$ indivually, obtain an orthonormal basis $\{c_1^i, \ldots, c_{n_i}^i\}$ of $E_{\lambda_i}(f)$ for each $i \in \{1, 2, \ldots, k\}$.
- Then $\bigcup_{i=1}^k \{c_1^i, \ldots, c_{n_i}^i\}$ is an orthonormal basis of K^n consisting of eigenvectors of f.

You will have the opportunity to practise this method on Sheet 9 of the exercises.

Chapter 11

Convergence and continuity in metric spaces

11.1 Metric spaces and convergence

In this chapter we want to prepare the development of multivariable calculus by establishing very general forms of convergence of sequences and continuity of maps on more general sets than just $\mathbb{R}$ and $\mathbb{C}$. in sets that generalize convergence in $\mathbb{R}$ in $\mathbb{C}$ as we have discussed in Chapter 2.
Most important notions from Mathematics 1, in particular convergence and continuity, have been defined in terms of numbers that *lie close* to one another. Continuity and differentiability of a function in a point are defined in terms of the behavior of the function *near that point*, i.e. can not be defined by considering the value of the function in that point alone. For this reason, we need to generalize the notion of points being *close to* or *nearby* one another. We will generalize is by introducting functions that measure *distances* between two points in a sense explained below.
In $\mathbb{R}$ and $\mathbb{C}$ we have often expressed that two points$x$ and y are close to one another by saying that $|x-y|<\varepsilon$, so closeness was formalized in terms of the absolute value.
Taking a look at the proofs of results about convergence and continuity, it turns out that we mostly used only the following elementary properties of the absolute value both in $K=\mathbb{R}$ and in $K=\mathbb{C}$:

- $|x|\geq 0$ for all $x\in K$ and $|x|=0$ if and only if $x=0$,
- $|x|=|-x|$ for all $x\in K$, in particular $|x-y|=|y-x|$ for all $x,y\in K$,
- $|x+y|\leq |x|+|y|$ for all $x,y\in K$.

These elementary properties alone already allow for a lot of computations and estimates. Taking a closer look at them, one realizes that by definition of a norm, the analogous properties are satisfied for norms on real or complex vector spaces. In fact, the notions of convergence and continuity from Chapters 2 and 3 can be generalized in a straightforward way to normed vector spaces by simply replacing the absolute value by the respective norm. In this section, we want to go one step further and introduce the notion of a metric space, i.e. sets with distance functions. We will see that normed vector spaces occur as special (and for our purposes the most important) cases of that construction.

Definition 11.1. Let X be a set. A map $d : X \times X \to \mathbb{R}$ is called *metric* or *distance function on* X if it has the following properties:

(i) (positive definiteness) $d(x,y) \geq 0$ for all $x, y \in X$ and $d(x,y) = 0$ holds if and only if $x = y$.

(ii) (symmetry) $d(x,y) = d(y,x)$ for all $x, y \in X$.

(iii) (triangle inequality) $d(x,y) \leq d(x,y) + d(y,z)$ for all $x, y, z \in X$.

If d is a metric on X, then the pair (X,d) is called a *metric space*. One simply writes X instead of (X,d) if it is evident which metric one is referring to.

As mentioned in the introduction of this section, norms on vector spaces can be used to define metrics on vector spaces.

Theorem/Definition 11.2. *Let $(V, \|\cdot\|)$ be a normed real or complex vector space. Then*

$$d : V \times V \to \mathbb{R}, \quad d(x,y) = \|x - y\|,$$

is a metric on V. We call d the metric induced by $\|\cdot\|$.

Proof. By property (i) of norms, it holds that $d(x,y) = \|x - y\| \geq 0$ for all $x, y \in V$ with $\|x - y\| = 0$ if and only if $x - y = 0$, i.e. $x = y$. Hence, d has property (i) of Definition 11.1.
Using property (ii) of norms, we compute for all $x, y \in V$ that

$$d(x,y) = \|x - y\| = \|-(y - x)\| = |-1| \cdot \|y - x\| = \|y - x\| = d(y,x).$$

Hence, d has property (ii) of Definition 11.1.
By the triangle inequality for norms, it further holds for all $x, y, z \in X$ that

$$d(x,y) = \|x - y\| = \|(x - z) + (z - y)\| \leq \|x - z\| + \|z - y\| = d(x,z) + d(z,y),$$

which shows the third property from Definition 11.1. Hence, d is a metric on V. □

Example 11.3. (i) The metric on $\mathbb{R}^n$ that is induced by the standard norm,

$$d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

is called the *Euclidean metric* or *standard metric* on $\mathbb{R}^n$.

(ii) On each set X there exists a metric given by

$$d_0 : X \times X \to \mathbb{R}, \qquad d_0(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

One checks that d_0 has properties (i), (ii) and (iii) from Definition 11.1, so d_0 is indeed a metric, called the *discrete metric*. (As one might guess, the discrete metric is not exactly useful for computations and applications...)

(iii) If (X,d) is a metric space and $A \subset X$, then the restriction $d_A := d|_{A \times A} : A \times A \to \mathbb{R}$ is a metric on A.

Next we want to introduce a general definition of convergence for arbitrary metric spaces. Keeping in mind that a metric on $\mathbb{R}$ or on $\mathbb{C}$ is defined by $d(x,y) = |x-y|$, the general definition is very plausible with the definition of convergence from Chapter 2 in mind.

Definition 11.4. Let (X,d) be a metric space, $(x_n)_{n\in\mathbb{N}}$ be a sequence in X and $a \in X$.. The sequence $(x_n)_{n\in\mathbb{N}}$ *converges to a* if

$$\forall \varepsilon > 0 \quad \exists n_0 \in \mathbb{N}: \quad d(x_n, a) < \varepsilon \; \forall n \geq n_0.$$

Then a is called the *limit* of $(x_n)_{n\in\mathbb{N}}$ and we denote it by

$$\lim_{n\to x_n} x_n = a.$$

A sequence in X is called *convergent* if it converges to some limit in X. A sequence in X that is not convergent is called *divergent*.

Remark 11.5. (1) Note that for a normed vector space $(V, \|\cdot\|)$ with the induced metric, the definition of convergence of a sequence $(x_n)_{n\in\mathbb{N}}$ in V resembles the one from Chapter 2 even more:

$$\forall \varepsilon > 0 \; \exists n_0 \in \mathbb{N} \; : \; \|x_n - a\| < \varepsilon \; \forall n \geq n_0.$$

(2) One checks from the definition of convergence that a sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space (X,d) converges to $a \in X$ if and only if the real sequence $(d(x_n,a))_{n\in\mathbb{N}}$ converges to 0 in $\mathbb{R}$ in the sense of Chapter 2.

The following theorem and its proof are line-by-line analogues of Theorem 2.4 and its proof.

Theorem 11.6. *Let (X,d) be a metric space. The limit of a convergent sequence in X is unique.*

Proof. (This proof is not discussed in the lecture videos.)

Let $a, b \in X$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X which converges both to a and to b. Let $\varepsilon > 0$ be arbitrary. By assumption there exist $n_1, n_2 \in \mathbb{N}$, such that

$$\begin{aligned} d(x_n, a) &< \varepsilon \quad \forall n \geq n_1, \\ d(x_n, b) &< \varepsilon \quad \forall n \geq n_2. \end{aligned}$$

If we put $n_0 := \max\{n_1, n_2\}$, then we obtain for each $n \geq n_0$ using the triangle inequality and the symmetry of d that

$$d(a,b) \leq d(a,x_n) + d(x_n,b) = d(x_n,a) + d(x_n,b) < \varepsilon + \varepsilon = 2\varepsilon.$$

So we have shown that $d(a,b) < 2\varepsilon$ for all $\varepsilon > 0$. If we had $a \neq b$, then by positive definiteness of a metric $d(a,b) > 0$, so choosing $\varepsilon = \frac{d(a,b)}{2}$ would yield $d(a,b) < d(a,b)$, which is a contradiction. Hence, $a = b$, which shows the claim. □

The following theorem shows that convergence of a sequence in $\mathbb{R}^n$ with respect to the Euclidean metric is equivalent to the convergence of all its component sequences. This reconciles the formal definition with our intuitive notion about convergence in $\mathbb{R}^n$.

Theorem 11.7. *Let $k \in \mathbb{N}$ and consider $\mathbb{R}^k$ with the Euclidean metric. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^k$, where $x_n = (x_{1n}, x_{2n}, \ldots, x_{kn})$ for each $n \in \mathbb{N}$, and let $a = (a_1, \ldots, a_k) \in \mathbb{R}^k$. Then $(x_n)_{n\in\mathbb{N}}$ converges to a if and only if $(x_{in})_{n\in\mathbb{N}}$ converges to a_i in $\mathbb{R}$ for all $i \in \{1, 2, \ldots, k\}$.*

Proof. "⇒": We observe from the monotonicity of roots that for all $i \in \{1, 2, \ldots, k\}$ and $n \in \mathbb{N}$ it holds that

$$|x_{in} - a_i| = \sqrt{(x_{in} - a_i)^2} \leq \sqrt{\sum_{j=1}^{k} (x_{jn} - a_j)^2} = \|x_n - a\| = d(x_n, a). \tag{11.1}$$

Let $\varepsilon > 0$. By assumption, there exists $n_0 \in \mathbb{N}$, such that $d(x_n, a) < \varepsilon$ for all $n \geq n_0$. By (11.1) this implies for each $i \in \{1, 2, \ldots, k\}$ that

$$|x_{in} - a_i| \leq d(x_n, a) < \varepsilon \qquad \forall n \geq n_0.$$

Since ε is arbitrary, this shows that $\lim_{n\to\infty} x_{in} = a_i$ for each $i \in \{1, 2, \ldots, k\}$.
"⇐": Let $\varepsilon > 0$. By assumption for each $i \in \{1, 2, \ldots, k\}$ there exists $n_i \in \mathbb{N}$, such that

$$|x_{in} - a_i| < \frac{\varepsilon}{\sqrt{k}} \qquad \forall n \geq n_i.$$

Put $n_0 := \max\{n_1, \ldots, n_k\}$. Then for $n \geq n_0$ we compute, using the strict monotonicity of roots:

$$d(x_n, a) = \|x_n - a\| = \sqrt{\sum_{i=1}^{k} (x_{in} - a_i)^2} < \sqrt{\sum_{i=1}^{k} \frac{\varepsilon^2}{k}} = \sqrt{k \cdot \frac{\varepsilon^2}{k}} = \sqrt{\varepsilon^2} = \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that $(x_n)_{n\in\mathbb{N}}$ converges to a. □

While sums and multiples of sequences do not make any sense in arbitrary metric spaces, these notions of course exist in vector spaces and it turns out that convergence in normed vector spaces behaves analogously to convergence in $\mathbb{R}$ or $\mathbb{C}$. Before proceeding we fix a notational convention:

In the following, unless otherwise mentioned, we consider a normed vector space as equipped with the metric induced by its norm.

Theorem 11.8. *Let $(V, \|\cdot\|)$ be a normed vector space over $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ be convergent sequences in V. Let $a = \lim_{n\to\infty} x_n$ and $b = \lim_{n\to\infty} y_n$. Then:*

a) $(x_n + y_n)_{n\in\mathbb{N}}$ converges against $a + b$,

b) $(\lambda \cdot x_n)_{n\in\mathbb{N}}$ converges against λa for all $\lambda \in K$.

Proof. This is shown along the same lines as Theorem 2.7 with absolute values replaced by $\|\cdot\|$ everywhere. □

The following definition generalizes another basic notion for sequences from real or complex numbers to arbitrary metric spaces.

Definition 11.9. Let (X, d) be a metric space and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. We call $(x_n)_{n\in\mathbb{N}}$ *bounded* if for an $x_0 \in X$ there exists $R > 0$, such that $d(x_n, x_0) \leq R$ for all $n \in \mathbb{N}$.

The following result is the immediate generalization of Theorem 2.10. One checks that it is proven along the same lines.

Theorem 11.10. *Let (X, d) be a metric space. Every convergent sequence in X is bounded.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X with $\lim_{n\to\infty} x_n = a \in X$. In particular, there exists $n_0 \in \mathbb{N}$ with

$$d(x_n, a) < 1 \quad \forall n \geq n_0.$$

Thus, if we put $R := \max\{d(x_1, a), d(x_2, a), \ldots, d(x_{n_0-1}, a), 1\}$, then $d(x_n, a) \leq R$ for all $n \in \mathbb{N}$, which shows that the sequence is bounded. □

The following definition generalizes the notion of Cauchy sequences in $\mathbb{R}$ and $\mathbb{C}$ from Definition 2.50.

Definition 11.11. Let (X, d) be a metric space.

a) A sequence $(x_n)_{n\in\mathbb{N}}$ in X is a *Cauchy sequence* if

$$\forall \varepsilon > 0 \; \exists n_0 \in \mathbb{N} : \quad d(x_n, x_m) < \varepsilon \;\; \forall m, n \geq n_0.$$

b) (X, d) is *complete* if every Cauchy sequence in X converges.

Example 11.12. (1) By Theorem 2.51 and Corollary 2.52, $\mathbb{R}$ and $\mathbb{C}$ with the Euclidean metrics (i.e. the distance function defined by the absolute value) is complete in the sense of Definition 11.11.

(2) $(0, 1)$ with the metric induced by the Euclidean metric on $\mathbb{R}$ is not complete. For example, the sequence $(\frac{1}{n})_{n\in\mathbb{N}}$ is a Cauchy sequence, but it does not converge in $(0, 1)$ since $0 \notin (0, 1)$, which is its limit in $\mathbb{R}$.

Theorem 11.13. *Let (X, d) be a metric space. Then every convergent sequence in X is a Cauchy sequence.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X which converges to $a \in X$. Let $\varepsilon > 0$. Then there exists $n_0 \in \mathbb{N}$, such that $d(x_n, a) < \frac{\varepsilon}{2}$ for all $n \geq n_0$. Then by the triangle inequality and symmetry of d

$$d(x_n, x_m) \leq d(x_n, a) + d(a, x_m) = d(x_n, a) + d(x_m, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad \forall m, n \geq n_0.$$

Thus, $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. □

Remark 11.14. While the opposite implication in Theorem 11.13 is true for $X = \mathbb{R}$ by Theorem 2.51, this is no longer the case for arbitrary metric spaces. Taking a look at the proof of Theorem 2.51, one observes that the proof of the opposite implication made use of the Bolzano-Weierstraß theorem which does not hold true for arbitrary metric spaces (we will make this precise later), so the proof can not be transferred to the present setting.j

Theorem 11.15. *Let $k \in \mathbb{N}$ and let $\mathbb{R}^k$ be equipped with the Euclidean metric. Then $\mathbb{R}^k$ is complete.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$, where $x_n = (x_{1n}, \ldots, x_{kn})$ for each $n \in \mathbb{N}$, be a Cauchy sequence in $\mathbb{R}^k$. Let $\varepsilon > 0$. Then there exists $n_0 \in \mathbb{N}$ with $\|x_n - x_m\| = d(x_n, x_m) < \varepsilon$ for all $n, m \geq n_0$. Thus, for each $i \in \{1, 2, \ldots, k\}$ we compute that

$$|x_{in} - x_{im}| \leq \|x_n - x_m\| < \varepsilon \quad \forall m, n \geq n_0.$$

Since $\varepsilon > 0$ was arbitrary, $(x_{in})_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$ for each i. By Theorem 2.51, $(x_{in})_{n\in\mathbb{N}}$ is convergent for each $i \in \{1, 2, \ldots, k\}$ and it follows from Theorem 11.7 that $(x_n)_{n\in\mathbb{N}}$ converges. Hence, $\mathbb{R}^k$ is complete. □

For normed vector spaces there is another common term used in the case of completeness.

Definition 11.16. A normed vector space $(V, \|\cdot\|)$ that is complete with respect to the metric induced by $\|\cdot\|$ is called a *Banach space*.

Banach spaces are one of the main objects of study in functional analysis and will appear in later lectures whenever one studies spaces of functions. In the following theorem, we recall the sup norm from Theorem/Definition 10.13, which was our first example of a norm on a vector space consisting of functions.

Theorem 11.17. *Let X be a set and let $K = \mathbb{R}$ or $K = \mathbb{C}$. Then $(\mathcal{B}(X,K), \|\cdot\|_\infty)$ is a Banach space.*

Proof. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{B}(X,K)$. Then for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $\sup_{x\in X} |f_n(x) - f_m(x)| < \varepsilon$ for all $n, m \geq n_0$. In particular, for each fixed $x \in X$ we obtain $|f_n(x) - f_m(x)| < \varepsilon$ for all $n, m \geq n_0$, which means that $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence in K for each $x \in X$. By the completeness of $\mathbb{R}$ and $\mathbb{C}$, i.e. by Theorem 2.51 and Corollary 2.52, this yields that $(f_n(x))_{n\in\mathbb{N}}$ converges in K for each $x \in X$. We define a function

$$f : X \to K, \quad f(x) := \lim_{n\to\infty} f_n(x).$$

We want to show that f is bounded and that $(f_n)_{n\in\mathbb{N}}$ converges to f. To show the latter, let $\varepsilon > 0$ and let $n_0 \in \mathbb{N}$, such that $\|f_n - f_m\|_\infty < \frac{\varepsilon}{2}$ for all $m, n \geq n_0$. Then for each $n \geq n_0$ and $x \in X$ we compute

$$\begin{aligned} |f_n(x) - f(x)| &\leq |f_n(x) - f_m(x)| + |f_m(x) - f(x)| \\ &\leq \|f_n - f_m\|_\infty + |f_m(x) - f(x)| < \frac{\varepsilon}{2} + |f_m(x) - f(x)|. \end{aligned}$$

Since this holds for every $m \geq n_0$ and since $\lim_{m\to\infty} |f_m(x) - f(x)| = 0$, it follows that

$$|f_n(x) - f(x)| \leq \frac{\varepsilon}{2} \qquad \forall n \geq n_0, x \in X,$$

hence

$$\|f_n - f\|_\infty = \sup_{x\in X} |f_n(x) - f(x)| \leq \frac{\varepsilon}{2} < \varepsilon \quad \forall n \geq n_0.$$

Since $\varepsilon > 0$ was arbitrary, this shows that $(f_n)_{n\in\mathbb{N}}$ converges to f. To see that f is bounded, we let $n \in \mathbb{N}$, such that $\|f - f_n\|_\infty < 1$. Then by the triangle inequality, we obtain for all $x \in X$:

$$|f(x)| \leq \sup_{x\in X} |f(x)| = \|f\|_\infty \leq \|f - f_n\|_\infty + \|f_n\|_\infty \leq 1 + \|f_n\|_\infty.$$

Since f_n is bounded, this shows that f is bounded as well, hence $f \in \mathcal{B}(X,K)$. Thus, $\mathcal{B}(X,K)$ is complete. □

11.2 Uniform convergence of sequences of functions

In Theorem 11.17 we have used our new general definition of convergence to study convergence of sequences of functions. In this section we want to further delve into notions of convergence for sequences of functions. For the sake of simplicity, we will restrict

Definition 11.18. Let X be a set and let Y be a metric space. Let $f_n : X \to Y$, $n \in \mathbb{N}$, and $f : X \to Y$ be maps. We say that $(f_n)_{n\in\mathbb{N}}$ *converges pointwise to* f if

$$\lim_{n\to\infty} f_n(x) = f(x) \qquad \forall x \in X.$$

f is then called the *pointwise limit of* $(f_n)_{n\in\mathbb{N}}$.

The notion of pointwise convergence is an elementary and simple definition of convergence, but is it a good one? Let us consider the familiar case of $Y = \mathbb{R}$ and $X = I$ an interval for a moment. We would like a notion of limit of a sequence of function, for which the limit function *inherits* certain properties from the sequence. More precisely, we would like to have positive answers to the following questions, where $(f_n)_{n\in\mathbb{N}}$ always converges pointwise to f.

(1) If $f_n : I \to \mathbb{R}$ is continuous for each $n \in \mathbb{N}$, is f again continuous?

(2) If $f_n : I \to \mathbb{R}$ is differentiable for each $n \in \mathbb{N}$, is f again differentiable and does $(f_n')_{n\in\mathbb{N}}$ converge against f'?

(3) If $f_n : [a, b] \to \mathbb{R}$ is Riemann-integrable for each $n \in \mathbb{N}$, is f again Riemann-integrable and is $\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx$?

Unfortunately, the answer to each of the three questions is No, as the following counterexamples show.

(1) For each $n \in \mathbb{N}$ consider the continuous function $f_n : [0, 1] \to \mathbb{R}$, $f_n(x) = x^n$. Then $\lim_{n\to\infty} f_n(x) = f(x)$ for each $x \in [0, 1]$, where

$$f : [0, 1] \to \mathbb{R}, \quad f(x) = \begin{cases} 0 & \text{if } x \in [0, 1), \\ 1 & \text{if } x = 1. \end{cases}$$

which is not continuous in 1.

(2) For each $n \in \mathbb{N}$ let $f_n : [0, 1] \to \mathbb{R}$ be the function whose graph is shown in Figure 11.1. Then

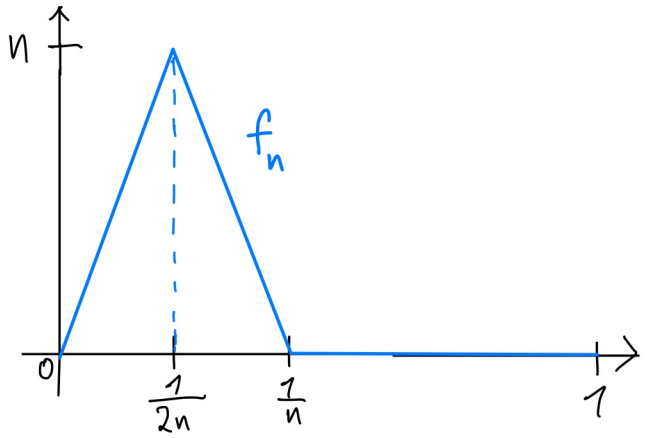

Figure 11.1: The graph of f_n in the third counterexample.

$\lim_{n\to\infty} f_n(x) = 0$ for all $x \in [0,1]$ since for each $x > 0$ it holds that $f_n(x) = 0$ whenever $n > \frac{1}{x}$ and since $f_n(0) = 0$ for all $n \in \mathbb{N}$. Hence, $(f_n)_{n\in\mathbb{N}}$ converges pointwise against the zero function, but

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,dx = \lim_{n\to\infty} \frac{1}{2} = \frac{1}{2} \neq 0 = \int_0^1 0\,dx.$$

(3) For each $n \in \mathbb{N}$ let $f_n : [0, 2\pi] \to \mathbb{R}$, $f_n(x) = \frac{\sin(nx)}{\sqrt{n}}$, which is differentiable for each $n \in \mathbb{N}$. Then

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{\sin(nx)}{\sqrt{n}} = 0$$

for each $x \in [0, 2\pi]$, so $(f_n)_{n\in\mathbb{N}}$ converges pointwise against the zero function. This function is differentiable, but by the chain rule

$$f_n'(x) = \frac{n\cos(nx)}{\sqrt{n}} = \sqrt{n}\cos(nx).$$

Hence, $(f_n')_{n\in\mathbb{N}}$ is divergent.

This shows that while pointwise convergence is a very elementary and simple notion of convergence for sequences of functions, it is not very practical for computations. For this purpose, we introduce another notion of convergence that is defined in terms of a norm.

Definition 11.19. Let X be a set, let $K = \mathbb{R}$ or $K = \mathbb{C}$, let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{B}(X, K)$ and let $f \in \mathcal{B}(X, K)$. We say that $(f_n)_{n\in\mathbb{N}}$ *converges uniformly* to f if $(f_n)_{n\in\mathbb{N}}$ converges to f with respect to $\|\cdot\|_\infty$, i.e. if

$$\lim_{n\to\infty} \|f_n - f\|_\infty = 0.$$

In this case, we write $\lim_{n\to\infty} f_n = f$ and call f the *uniform limit* or just *limit* of $(f_n)_{n\in\mathbb{N}}$.

Example 11.20. (1) Let $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = (x^2+1)e^{x/n}$. We compute the pointwise limit of f_n as

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} (x^2+1)e^{x/n} = (x^2+1)\lim_{n\to\infty} e^{x/n} = (x^2+1)e^0 = x^2+1.$$

We want to show that $(f_n)_{n\in\mathbb{N}}$ indeed converges uniformly to $f(x) = x^2+1$. For each $x \in [0,1]$ it holds that

$$|f_n(x) - f(x)| = |(x^2+1)(e^{x/n}-1)| = |x^2+1|\cdot|e^{x/n}-1| \leq 2|e^{x/n}-1|.$$

By strict monotonicity of the exponential function, $1 < e^{x/n} \leq e^{1/n}$, so it follows that

$$|f_n(x) - f(x)| \leq 2(e^{1/n}-1) \quad \Rightarrow \quad \|f_n - f\|_\infty = \sup_{x\in[0,1]} |f_n(x) - f(x)| \leq 2(e^{1/n}-1).$$

By continuity of the exponential function, $\lim_{n\to\infty}(e^{1/n}-1) = 0$, hence for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ with $e^{1/n} - 1 < \varepsilon$ for all $n \geq n_0$ and thus

$$\|f_n - f\|_\infty < \varepsilon \quad \forall n \geq n_0.$$

Hence, $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f.

(2) In the counterexamples to questions (1) and (2) the sequences $(f_n)_{n\in\mathbb{N}}$ do not converge uniformly against their pointwise limits. (It is highly recommended to the reader to check this by themselves!) Thus, the converse implication does not hold, there are pointwise convergent sequences that do not converge uniformly.

The following statement characterizes uniform convergence more explicitly without making use of the sup norm.

Theorem 11.21. *Let X be a set, $K = \mathbb{R}$ or $K = \mathbb{C}$, $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{B}(X,K)$ and $f \in \mathcal{B}(X,K)$. The following statements are equivalent:*

(i) $(f_n)_{n\in\mathbb{N}}$ *converges uniformly to* f.

(ii) $\forall \varepsilon > 0\ \exists n_0 \in \mathbb{N}: \quad |f_n(x) - f(x)| < \varepsilon \quad \forall n \geq n_0, x \in X.$

Proof. (i) $\Rightarrow$ (ii): If $\lim_{n\to\infty} \|f_n - f\| = 0$, then for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$, such that

$$\|f_n - f\|_\infty < \varepsilon \quad \forall n \geq n_0 \quad \Leftrightarrow \quad \sup_{x\in X} |f_n(x) - f(x)| < \varepsilon \quad \forall n \geq n_0.$$

By definition of the supremum, this yields

$$|f_n(x) - f(x)| < \varepsilon \quad \forall n \geq n_0,\ x \in X,$$

which shows the claim.
(ii) $\Rightarrow$ (i): If (ii) holds, then for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ so that

$$\|f_n - f\|_\infty = \sup_{x\in X} |f_n(x) - f(x)| \leq \varepsilon \quad \forall n \geq n_0.$$

This yields that $\lim_{n\to\infty} \|f_n - f\|_\infty = 0$, which we wanted to show. □

Remark 11.22. Let X be a set, let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{B}(X,K)$.

(1) One derives from Theorem 11.21 that if $(f_n)_{n\in\mathbb{N}}$ converges uniformly against $f \in \mathcal{B}(X,K)$, then $(f_n)_{n\in\mathbb{N}}$ converges pointwise against f.

(2) In practice, to check whether a given sequence of functions converges uniformly, one already needs a "candidate" for the limit of the sequence. Such a limit is usually obtained by *explicitly computing the pointwise limit of the sequence*, which is much easier to determine, and *then checking whether the convergence is uniform or not*. If the sequence converges pointwise then by (1) the only possibility for uniform convergence is convergence to the pointwise limit.

(3) Given Theorem 11.21 we can spot the difference between pointwise and uniform convergence more explicitly. If a sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to f, it holds for fixed x and given $\varepsilon > 0$ that there exists an $n_0 \in \mathbb{N}$ *which depends on ε and x* that $|f_n(x) - f(x)| < \varepsilon$ for $n \geq n_0$. If the convergence is uniform, then *the choice of n_0 no longer depends on x, but must be the same for all x in the domain.*

We will next show that uniform convergence does not have the same "flaws" as pointwise convergence when it comes to the questions from the beginning of the section.

Theorem 11.23. *Let $D \subset \mathbb{C}$ and let $K = \mathbb{R}$ or $K = \mathbb{C}$. Let $f_n : D \to K$ be bounded for each $n \in \mathbb{N}$. If f_n is continuous for each $n \in \mathbb{N}$ and if $(f_n)_{n\in\mathbb{N}}$ converges uniformly to $f : D \to K$, then f is continuous.*

Proof. Let $x_0 \in D$. We want to show that f is continuous in x_0, i.e. that for all $\varepsilon > 0$ there exists $\delta > 0$, such that $|f(x) - f(x_0)| < \varepsilon$ for all $x \in D$ with $|x - x_0| < \delta$.
Let $\varepsilon > 0$. Since $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f, there exists $n_0 \in \mathbb{N}$ with $\|f_{n_0} - f\|_\infty < \frac{\varepsilon}{3}$ and thus

$$|f_{n_0}(x) - f(x)| < \frac{\varepsilon}{3} \qquad \forall x \in D. \tag{11.2}$$

Since f_{n_0} is continuous, there exists $\delta > 0$, such that

$$|f_{n_0}(x) - f_{n_0}(x_0)| < \frac{\varepsilon}{3} \quad \forall x \in D \text{ with } |x - x_0| < \delta. \tag{11.3}$$

Hence, we derive for all $x \in D$ with $|x - x_0| < \delta$ that

$$\begin{aligned} |f(x) - f(x_0)| &= |f(x) - f_{n_0}(x) + f_{n_0}(x) - f_{n_0}(x_0) + f_{n_0}(x_0) - f(x_0)| \\ &\leq |f_{n_0}(x) - f(x)| + |f_{n_0}(x) - f_{n_0}(x_0)| + |f_{n_0}(x_0) - f(x_0)| \\ &\overset{(11.2),(11.3)}{<} \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Since ε was chosen arbitrarily, f is continuous in x_0. Since x_0 was chosen arbitrarily, f is continuous. □

Corollary 11.24. *Let $D \subset \mathbb{C}$ and let*

$$C_b^0(D, K) := \{f \in C^0(D, K) \mid f \text{ is bounded}\}.$$

Then $(C_b^0(D, K), \|\cdot\|_b)$ is a Banach space, where $\|\cdot\|_b$ denotes the restriction of the sup norm to $C_b^0(D, K)$.

Proof. (This proof is not discussed in the lecture videos.)

As sums and scalar multiples of continuous and bounded functions are continuous and bounded, $C_b^0(D, K)$ is a linear subspace of $\mathcal{B}(D, K)$. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $C_b^0(D, K)$. By definition of the norm, $(f_n)_{n\in\mathbb{N}}$ is also a Cauchy sequence in $\mathcal{B}(D, K)$, thus, by Theorem 11.17, converges uniformly against some $f \in \mathcal{B}(D, K)$. By Theorem 11.23, f is again continuous, hence $f \in C_b^0(D, K)$. Thus, $(f_n)_{n\in\mathbb{N}}$ converges against f in $C_b^0(D, K)$ as well. □

Theorem 11.25. *Let $a, b \in \mathbb{R}$ with $a < b$, let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{B}([a, b], \mathbb{R})$ and let $f : [a, b] \to \mathbb{R}$ be bounded. If f_n is Riemann-integrable for every $n \in \mathbb{N}$ and if $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f, then f is Riemann-integrable with*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx.$$

Proof. (The proof of the Riemann-integrability of f is not discussed in the lecture videos.[1])

We first need to show that f is Riemann-integrable. Let $\varepsilon > 0$. Since $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f, there exists $n_0 \in \mathbb{N}$ with

$$\|f_{n_0} - f\|_\infty < \frac{\varepsilon}{3(b - a)}.$$

[1] If one makes the additional assumption that the f_n are continuous, then the argument fits within one line: By Theorem 11.23, f is again continuous, hence Riemann-integrable.

Then

$$f_{n_0}(x) - \frac{\varepsilon}{3(b-a)} < f(x) < f_{n_0}(x) + \frac{\varepsilon}{3(b-a)} \qquad \forall x \in [a,b].$$

Hence, if $J \subset [a,b]$ is an arbitrary interval with $x \in J$ it follows that

$$\inf_{t\in J} f_{n_0}(t) - \frac{\varepsilon}{3(b-a)} < f(x) < \sup_{t\in J} f_{n_0}(t) + \frac{\varepsilon}{3(b-a)}.$$

Consequently,

$$\sup_{t\in J} f(t) \le \sup_{t\in J} f_{n_0}(t) + \frac{\varepsilon}{3(b-a)}, \qquad \inf_{t\in J} f(t) \ge \inf_{t\in J} f_{n_0}(t) - \frac{\varepsilon}{3(b-a)}. \tag{11.4}$$

Thus, if $P = (t_0, t_1, \dots, t_r) \in \mathcal{P}([a,b])$ is a partition, then by the first inequality from (11.4)

$$\begin{aligned} U(f,P) &= \sum_{j=0}^{r-1} \sup_{t\in[t_j,t_{j+1}]} f(t)(t_{j+1}-t_j) \\ &\le \sum_{j=0}^{r-1} \Big(\sup_{t\in[t_j,t_{j+1}]} f_{n_0}(t) + \frac{\varepsilon}{3(b-a)} \Big)(t_{j+1}-t_j) \\ &= \sum_{j=0}^{r-1} \sup_{t\in[t_j,t_{j+1}]} f_{n_0}(t)(t_{j+1}-t_j) + \frac{\varepsilon}{3(b-a)}(b-a) \\ \Rightarrow U(f,P) &\le U(f_{n_0},P) + \frac{\varepsilon}{3} \end{aligned}$$

Completely analogously, the second inequality from (11.4) yields

$$L(f,P) \ge L(f_{n_0},P) - \frac{\varepsilon}{3}.$$

Since f_{n_0} is Riemann-integrable, there exists $P_0 \in \mathcal{P}([a,b])$ with $U(f_{n_0},P_0) - L(f_{n_0},P_0) < \frac{\varepsilon}{3}$. Combining this with the previous inequalities yields

$$U(f,P_0) - L(f,P_0) \le U(f_{n_0},P_0) + \frac{\varepsilon}{3} - L(f,P_0) + \frac{\varepsilon}{3} < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Hence, f is Riemann-integrable. *(In the lecture videos, we take it from here.)*

Using this observation and some of our results for integrals, more precisely Theorem 5.13.b) and Theorem 5.11.e), we further compute for all $n \in \mathbb{N}$ that

$$\left| \int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx \right| = \left| \int_a^b (f_n(x) - f(x))\,dx \right| \le \int_a^b |f_n(x) - f(x)|\,dx \le \|f_n - f\|_\infty (b-a).$$

Since $\lim_{n\to\infty} \|f_n - f\|_\infty = 0$ by assumption, this yields that $\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx$. □

The situation for differentiability is slightly more sophisticated than the one for integrability. In fact, there are (complicated) examples of uniformly convergent sequences of differentiable function, whose derivatives do not converge to the derivative of the uniform limit.

Theorem 11.26. *Let I be an interval and let $K = \mathbb{R}$ or $K = \mathbb{C}$. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{B}(I,K)$ and let $f : I \to K$ be bounded. Assume that*

- *f_n is continuously differentiable for each $n \in \mathbb{N}$,*
- *$(f_n)_{n\in\mathbb{N}}$ converges pointwise to f, and*
- *$(f_n')_{n\in\mathbb{N}}$ converges uniformly against $g : I \to K$.*

Then f is continuous differentiable with $f' = g$.

Proof. Since f_n' is continuous for each $n \in \mathbb{N}$ it follows from Theorem 11.23 that g is continuous. By the fundamental theorem of calculus (Theorem 5.19) it holds for all $n \in \mathbb{N}$ and $x \in [a, b]$ that

$$f_n(x) = f_n(a) + \int_a^x f_n'(t)\, dt.$$

Thus, by Theorem 11.25 and the pointwise convergence , we obtain for all $x \in [a, b]$ that

$$\begin{aligned} f(x) &= \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} f_n(a) + \lim_{n\to\infty} \int_a^x f_n'(t)\, dt \\ &= f(a) + \int_a^x g(t)\, dt. \end{aligned}$$

By part a) of the fundamental theorem of calculus, this implies that f is differentiable with $f'(x) = g(x)$ for all $x \in [a, b]$. □

11.3 Continuity of maps between metric spaces

In the same way as we transferred the notion of convergence to sequences in metric spaces, we will next generalize the notion of continuity to maps between arbitrary metric spaces. Again we will replace expressions in $\mathbb{R}$ or $\mathbb{C}$ of the form $|x - y|$ by distances $d(x, y)$ in metric spaces and it turns out that many of our previous results on continuity from Mathematics 1 are carried over to metric spaces and normed vector spaces along the same lines.

From this section on, unless otherwise mentioned, we always consider $\mathbb{R}^k$ as equipped with the Euclidean metric.

Definition 11.27. Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$ be a map.

a) Let $a \in X$. f is *continuous in a* if

$$\forall \varepsilon > 0\ \exists \delta > 0 : \quad d_Y(f(x), f(a)) < \varepsilon \quad \forall x \in X \text{ with } d_X(x, a) < \delta.$$

b) f is called *continuous* if f is continuous in a for every $a \in X$.

c) We put $C^0(X, Y) := \{f : X \to Y \mid f \text{ is continuous}\}$.

Remark 11.28. For normed vector spaces and their induced metrics, the definition of continuity looks even more like the one from Chapter 3: Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed vector spaces, $f : V \to W$ and $a \in V$. Then f is continuous in a if

$$\forall \varepsilon > 0\ \exists \delta > 0 : \quad \|f(x) - f(a)\|_W < \varepsilon \quad \forall x \in V \text{ with } \|x - a\| < \delta.$$

Example 11.29. (1) $\mathrm{id}_X : X \to X$ is continuous and constant maps $c_a : X \to X$, $c_a(x) = a$ for all $x \in X$, are continuous for every metric space (X, d). This follows straight from the definition of continuity.

(2) Let $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = (y,-x)$, is continuous: let $(x_0,y_0) \in \mathbb{R}^2$. Given $\varepsilon > 0$ put $\delta := \varepsilon > 0$. Then for all $(x,y) \in \mathbb{R}^2$ with $\|(x,y) - (x_0,y_0)\| < \delta$, it holds that

$$\begin{aligned}\|f(x,y) - f(x_0,y_0)\| &= \|(y,-x) - (y_0,-x_0)\| = \|(y-y_0,-x+x_0)\| \\ &= \sqrt{(y-y_0)^2 + (-x+x_0)^2} = \sqrt{(x-x_0)^2 + (y-y_0)^2} \\ &= \|(x,y) - (x_0,y_0)\| < \delta = \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, f is continuous in (x_0,y_0).

(3) Let $k \in \mathbb{N}$ and $i \in \{1,2,\dots,k\}$. Then $p_i : \mathbb{R}^n \to \mathbb{R}$, $p_i(x_1,\dots,x_n) = x_i$ is continuous: Let $a = (a_1,\dots,a_k) \in \mathbb{R}^k$. For a given $\varepsilon > 0$ put $\delta := \varepsilon$. Then for all $x = (x_1,\dots,x_k) \in \mathbb{R}^k$ with $\|x - a\| < \varepsilon$ we compute that

$$|p_i(x) - p_i(a)| = |x_i - a_i| \leq \|x - a\| < \delta = \varepsilon,$$

where $\|\cdot\|$ again denotes the standard norm on $\mathbb{R}^k$. Hence, p_i is continuous in a.

We will spend a great deal of this section by transferring results about continuous functions $D \to K$, where $D \subset K$ and $K = \mathbb{R}$ or $K = \mathbb{C}$ that we obtained in Chapter 3 to maps between metric spaces using our new definition of continuity. We will skip most of the proofs of the following theorems, as they are carried out line-by-line as the proofs of the corresponding results from Chapter 3. We begin with one of the most useful results, namely the characterization of continuity by convergent sequences.

Theorem 11.30 (Sequential continuity)**.** *Let (X,d_X) and (Y,d_Y) be metric spaces, let $f : X \to Y$ be a map and let $a \in X$. The following statements are equivalent:*

(i) f is continuous in a.

(ii) For every sequence $(x_n)_{n\in\mathbb{N}}$ in X with $\lim_{n\to\infty} x_n = a$ it holds that $\lim_{n\to\infty} f(x_n) = f(a)$.

Proof. This is shown along the same lines as Theorem 3.3, the corresponding result from Mathematics 1. □

Example 11.31. (1) We consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

Then f is continuous in $(0,0)$: for all $(x,y) \in \mathbb{R}^2 \setminus \{0\}$ we compute that

$$0 \leq |f(x,y)| = \left|\frac{x^2 y}{x^2+y^2}\right| = \frac{x^2|y|}{x^2+y^2} \leq \frac{(x^2+y^2)|y|}{x^2+y^2} = |y|.$$

Thus, if $(x_n,y_n)_{n\in\mathbb{N}}$ is any sequence in $\mathbb{R}^2$ with $\lim_{n\to\infty}(x_n,y_n) = (0,0)$, it holds that

$$0 \leq |f(x_n,y_n)| \leq |y_n| \qquad \forall n \in \mathbb{N}$$

which yields $\lim_{n\to\infty} |f(x_n,y_n)| = 0$ by the squeeze theorem. Hence,

$$\lim_{n\to\infty} f(x_n,y_n) = 0 = f(0,0)$$

for all $(x_n,y_n)_{n\in\mathbb{N}}$ converging to $(0,0)$. Theorem 11.30 then implies that f is continuous in $(0,0)$.

(2) For $a \in \mathbb{R}$ we consider the function

$$f_a : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ a & \text{if } (x,y) = (0,0). \end{cases}$$

Then f_a is *not* continuous in $(0,0)$ for any choice of a: If f_a was continuous in $(0,0)$, then by Theorem 11.30 it would hold that $\lim_{n\to\infty} f_a(x_n, y_n) = f_a(0,0) = a$ for all sequences $(x_n, y_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^2$ with $\lim_{n\to\infty}(x_n, y_n) = 0$. For $x_n = y_n = \frac{1}{n}$ we compute that

$$f_a\left(\tfrac{1}{n}, \tfrac{1}{n}\right) = \frac{\frac{1}{n^2}}{\frac{1}{n^2} + \frac{1}{n^2}} = \frac{1}{2}$$

for each $n \in \mathbb{N}$, hence $\lim_{n\to\infty} f_a(\frac{1}{n}, \frac{1}{n}) = \frac{1}{2}$. So f_a can only be continuous if $a = \frac{1}{2}$. But if we take $x_n = \frac{1}{n}$, $y_n = -\frac{1}{n}$, we likewise obtain

$$f_a\left(\tfrac{1}{n}, -\tfrac{1}{n}\right) = \frac{-\frac{1}{n^2}}{\frac{1}{n^2} + \frac{1}{n^2}} = -\frac{1}{2},$$

hence

$$\lim_{n\to\infty} f_a\left(\tfrac{1}{n}, -\tfrac{1}{n}\right) = -\tfrac{1}{2} \neq \tfrac{1}{2} = \lim_{n\to\infty} f_a\left(\tfrac{1}{n}, \tfrac{1}{n}\right).$$

So f_a is not continuous for any $a \in \mathbb{R}$.

Theorem 11.32. *Let (X, d_X), (Y, d_Y) and (Z, d_Z) be metric spaces, and let $a \in X$. Let $f : X \to Y$ be continuous in a and $g : Y \to Z$ be continuous in $f(a)$. Then $g \circ f : X \to Z$ is continuous in a.*

Proof. This follows from Theorem 11.30 in the same way that Theorem 3.7 was derived from Theorem 3.3. □

While sums and multiples by constants are not defined for arbitrary metric spaces, we can still generalize old results of sums and multiples by considering maps whose codomains are normed vector spaces. Here, sums and scalar multiples of maps are defined pointwise as usual and the corresponding result about continuity transfers to this setting.

Theorem 11.33. *Let (X, d) be a metric space, let E be a normed vector space over $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $f, g : X \to E$ be continuous in $a \in X$. Then:*

a) $f + g : X \to E$, $(f + g)(x) = f(x) + g(x)$, is continuous in a.

b) $\lambda f : X \to E$, $(\lambda f)(x) = \lambda \cdot f(x)$, is continuous in a for each $\lambda \in K$.

Proof. Both parts follow by combining Theorem 11.30 with the respective part of Theorem 11.8. □

In arbitrary normed vector spaces we still can not define products of functions as there is no multiplication of vectors inside of a normed vector space. However, if we further restrict to maps whose codomains are $\mathbb{R}$ or $\mathbb{C}$, we can again generalize the corresponding results from Chapter 3.

Theorem 11.34. *Let (X, d) be a metric space, let $K = \mathbb{R}$ or $K = \mathbb{C}$ and let $f, g : X \to K$ be continuous in $a \in X$. Then:*

a) $f \cdot g : X \to K$ *is continuous in a.*

b) If $g(x) \neq 0$ *for all* $x \in X$*, then* $\frac{f}{g} : X \to K$ *is continuous in a.*

Proof. This follows from Theorem 11.30 in the same way that Theorem 3.5 was derived from Theorem 3.3. □

One of the most important and most elementary classes of functions $K \to K$ are polynomials which we have studied in various contexts. The notion of a polynomial can be extended to maps in several real or complex variables, such that these polynomials share a lot of properties with single-variable ones. We next give a formal definition of a general polynomial and discuss their continuity.

Definition 11.35. Let $k \in \mathbb{N}$. A *(real) polynomial in k unknowns* is a map $p : \mathbb{R}^k \to \mathbb{R}$ for which there are $n_1, \ldots, n_k \in \mathbb{N}_0$, and $a_{i_1,\ldots,i_k} \in \mathbb{R}$ for all $i_j \in \{1, 2, \ldots, n_j\}$, $j \in \{1, 2, \ldots, k\}$, such that

$$p(x_1, x_2, \ldots, x_k) = \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \cdots \sum_{i_k=0}^{n_k} a_{i_1,i_2,\ldots,i_k} x_1^{i_1} x_2^{i_2} \ldots x_k^{i_k} \qquad \forall (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k.$$

The *degree of p* is given by

$$\deg p := \max\{i_1 + i_2 + \cdots + i_k \in \mathbb{N}_0 \mid a_{i_1,\ldots,i_k} \neq 0\}.$$

Example 11.36. (1) $p : \mathbb{R}^2 \to \mathbb{R}$, $p(x_1, x_2) = 1 + 2x_1 + x_2 + 2x_1^2 - 2x_1x_2^2$, is a polynomial in two unknowns of degree three.

(2) $p : \mathbb{R}^k \to \mathbb{R}$, $p(x_1, \ldots, x_k) = x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}$, is a polynomial in k variables of degree $n_1 + n_2 + \cdots + n_k$. A polynomial of this form is called *monomial*.

(3) A linear form $\varphi \in (\mathbb{R}^k)^*$, $\varphi(x_1, \ldots, x_k) = a_1x_1 + a_2x_2 + \cdots + a_kx_k$, is a polynomial of degree one.

Corollary 11.37. *Let* $k \in \mathbb{N}$*. Every polynomial* $p : \mathbb{R}^k \to \mathbb{R}$ *is continuous.*

Proof. Every polynomial is given by sums, products and scalar multiples of the $p_i : \mathbb{R}^k \to \mathbb{R}$ from Example 11.29.(2), which were shown to be continuous. Thus, the continuity of p follows from Theorems 11.33 and 11.34. □

Another notion which generalizes straightaway to metric spaces is Lipschitz continuity. Taking a look at Definition 3.8, it is evident how to generalize it in the spirit of the above definitions.

Definition 11.38. Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is *Lipschitz-continuous* if there exists $L \geq 0$, such that

$$d_Y(f(x_1), f(x_2)) \leq L \cdot d_X(x_1, x_2) \quad \forall x_1, x_2 \in X.$$

L is then called a *Lipschitz constant for* f.

Theorem 11.39. *Let* (X, d_X) *and* (Y, d_Y) *be metric spaces. Every Lipschitz-continuous map* $f : X \to Y$ *is continuous.*

Proof. This is shown along the same lines as Theorem 3.10. □

Example 11.40. Let (X,d) be a metric space, let $x_0 \in X$ and let $f : X \to \mathbb{R}$, $f(x) = d(x,x_0)$. For all $x,y \in X$ we compute that $d(x,x_0) \leq d(x,y) + d(y,x_0)$ and $d(y,x_0) \leq d(x,y) + (x,x_0)$, i.e. $f(x) - f(y) \leq d(x,y)$ and $f(y) - f(x) \leq d(x,y)$. Hence,

$$|f(x) - f(y)| \leq d(x,y) \quad \forall x,y \in X,$$

so f is Lipschitz-continuous with Lipschitz constant 1, hence f is continuous by Theorem 11.39.

We have seen in Theorem 11.7 that a sequence in $\mathbb{R}^k$ converges if and only if all of its component sequences converge. Using sequential continuity, we can derive a similar result for continuous maps from a metric space to $\mathbb{R}^k$.

Theorem 11.41. *Let (X,d) be a metric space, let $k \in \mathbb{N}$ and $f : X \to \mathbb{R}^k$, $f = (f_1, \ldots, f_k)$, where $f_i : X \to \mathbb{R}$ for each $i \in \{1,2,\ldots,k\}$. Let $a \in X$. Then f is continuous if and only if f_i is continuous for every $i \in \{1,2,\ldots,k\}$.*

Proof. We obtain the following chain of equivalences:

$$\begin{aligned}
& f \text{ is continuous in } a \\
\overset{\text{Theorem 11.30}}{\Leftrightarrow} \; & \lim_{n\to\infty} f(x_n) = f(a) \quad \text{for all } (x_n)_{n\in\mathbb{N}} \text{ in } X \text{ with } \lim_{n\to\infty} x_n = a \\
\overset{\text{Theorem 11.7}}{\Leftrightarrow} \; & \lim_{n\to\infty} f_i(x_n) = f_i(a) \quad \text{for all } i \in \{1,2,\ldots,k\} \text{ and } (x_n)_{n\in\mathbb{N}} \text{ in } X \text{ with } \lim_{n\to\infty} x_n = a \\
\overset{\text{Theorem 11.30}}{\Leftrightarrow} \; & f_i \text{ is continuous in } a \text{ for all } i \in \{1,2,\ldots,k\}.
\end{aligned}$$

□

Corollary 11.42. *Let $m,n \in \mathbb{N}$. Every linear map $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous.*

Proof. We have seen earlier that every linear map $f : \mathbb{R}^n \to \mathbb{R}^m$ is of the form $f = (f_1, \ldots, f_m)$, where $f_1, \ldots, f_m \in (\mathbb{R}^n)^*$. By Example 11.36.(3), the f_i are polynomials, hence continuous by Corollary 11.37. Thus, f is continuous by Theorem 11.41. □

Corollary 11.42 leads to the question if linear maps between normed vector spaces are always continuous. As linear maps are particularly "well-behaved" with respect to vector space structures, this might be a plausible assumption upon first thought. However, the following counterexample shows that this is actually *not* the case.

Example 11.43. Consider $C^1([0,1],\mathbb{R})$ with the restriction of the sup norm and consider the linear map $D : C^1([0,1],\mathbb{R}) \to C^0(\mathbb{R},\mathbb{R})$, $D(f) = f'$. Let $(f_n)_{n\in\mathbb{N}}$ be the sequence in $C^1([0,1],\mathbb{R})$ which is given as in counterexample (ii) in Section 4.2., i.e. $f_n(x) = \frac{\sin(nx)}{\sqrt{n}}$. Then

$$\|f_n\|_\infty = \sup_{x\in[0,2\pi]} \frac{|\sin(nx)|}{\sqrt{n}} = \frac{1}{\sqrt{n}} \qquad \forall n \in \mathbb{N},$$

hence $\lim_{n\to\infty} \|f_n\|_\infty = 0$, so $(f_n)_{n\in\mathbb{N}}$ converges to the zero function in $(C^1([0,2\pi],\mathbb{R}), \|\cdot\|_\infty)$. If D was continuous, then $\lim_{n\to\infty} D(f_n) \overset{!}{=} D(0) = 0$. But since $f_n'(x) = \sqrt{n}\cos(nx)$ we observe for each $n \in \mathbb{N}$ that

$$\|D(f_n)\|_\infty = \|f_n'\|_\infty = \sqrt{n} \sup_{x\in[0,2\pi]} |\cos(nx)| = \sqrt{n}.$$

In particular, $(\|D(f_n)\|_\infty)_{n\in\mathbb{N}}$ does not converge to 0, so $(D(f_n))_{n\in\mathbb{N}}$ does not converge to the zero function. Hence, D is not continuous.

Luckily, there is a criterion by which one can in most situation decide if a linear map is continuous or not.

Theorem 11.44. *Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed vector spaces and let $f : V \to W$ be linear. The following statements are equivalent:*

(i) f is continuous.

(ii) f is continuous in 0.

(iii) There exists $C \geq 0$, such that $\|f(v)\|_W \leq C \cdot \|v\|_V$ for all $v \in V$.

Proof. (i) $\Rightarrow$ (ii): This is trivial.
(ii) $\Rightarrow$ (iii): Since f is linear, $f(0) = 0$. Thus, if f is continuous in 0, then in particular (with $\varepsilon = 1$) there exists $\delta > 0$, such that

$$\|f(x)\|_W < 1 \quad \forall x \in V \text{ with } \|v\|_V < \delta.$$

Given $v \in V \setminus 0$, put $x_v := \frac{\delta}{2\|v\|} v$. Then $\|x_v\|_V = \frac{\delta}{2} < \delta$ and thus $\|f(x_v)\| < 1$. By the linearity of f and the properties of a norm, we derive

$$\begin{aligned} \|f(x_v)\|_W < 1 \quad &\Leftrightarrow \quad \left\| f\left(\frac{\delta}{2\|v\|} v\right) \right\|_W < 1 \quad \Leftrightarrow \quad \left\| \frac{\delta}{2\|v\|} f(v) \right\|_W < 1 \quad \Leftrightarrow \quad \frac{\delta}{2\|v\|} \|f(v)\|_W < 1 \\ &\Leftrightarrow \quad \|f(v)\|_W < \frac{2}{\delta} \|v\|_V. \end{aligned}$$

Thus, the claim follows with $C = \frac{2}{\delta}$.
(iii) $\Rightarrow$ (i): Since f is linear, we obtain by assumption that

$$\|f(v_1) - f(v_2)\|_W = \|f(v_1 - v_2)\|_W \leq C \cdot \|v_1 - v_2\|_V \qquad \forall v_1, v_2 \in V.$$

Hence, f is Lipschitz-continuous and thus continuous by Theorem 11.39. □

Example 11.45. Let I be an interval, $K = \mathbb{R}$ or $K = \mathbb{C}$ and consider the Banach space $(\mathcal{B}(I, K), \|\cdot\|_\infty)$. Let $h \in \mathcal{B}(I, K)$ and consider the map

$$L : \mathcal{B}(I, K) \to \mathcal{B}(I, K), \qquad L(f) = h \cdot f,$$

i.e. $(L(f))(x) = h(x)f(x)$ for each $x \in I$. Then L_f is linear. We compute for each $x \in I$ that

$$|(L(f))(x)| = |h(x)f(x)| \leq \|h\|_\infty \cdot |f(x)|,$$

hence with $C := \|h\|_\infty$ we obtain

$$\|L(f)\|_\infty \leq C\|f\|_\infty \qquad \forall f \in \mathcal{B}(I, K)$$

By Theorem 11.44, this shows that L is continuous.

11.4 Open and closed subsets of metric spaces

Before we turn towards higher-dimensional notions of differentiability, we will introduce subsets of metric spaces with useful properties, namely open subset and compact subsets.
In Chapter 4 of this course, we have studied the differentiability of functions defined on open intervals. An open interval I has the crucial property that for each $x_0 \in I$ there exists $\varepsilon > 0$, such that $(x_0 - \varepsilon, x_0 + \varepsilon) \subset I$. Intuitively there is "a little space" around x_0 inside of I. This is required, since the derivative is given by $f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$ and the computation of the limit requires the knowledge of the values of f near x_0.
The notion of an open subset of a metric space generalizes this idea. Intuitively, a subset is open, if there is "a little space" around every point inside the subset. The next definition formalizes this idea and introduces some additional related terminology.

Definition 11.46. Let (X, d) be a metric space.

a) For $x_0 \in X$ and $\varepsilon > 0$ we call

$$B_\varepsilon(x_0) := \{x \in X \mid d(x, x_0) < \varepsilon\}$$

the *open ε-ball centered in* x_0 or *around* x_0.

b) Let $x_0 \in X$. A subset $U \subset X$ is called a *neighborhood of* x_0 if there exists some $\varepsilon > 0$ with $B_\varepsilon(x_0) \subset U$.

c) A subset $U \subset X$ is *open in* X if for each $x_0 \in U$ there exists $\varepsilon > 0$ with $B_\varepsilon(x_0) \subset U$, i.e. if U is a neighborhood of each of its points.

Example 11.47. (1) Every open interval (a, b), where $a \in \mathbb{R} \cup \{-\infty\}$ and $b \in \mathbb{R} \cup \{+\infty\}$ is open in $\mathbb{R}$:

First of all note that for all $x \in \mathbb{R}$ and $r > 0$ it holds that $B_r(x) = (x - r, x + r)$. Thus, if for $x \in (a, b)$ we put

$$\varepsilon := \min\{x - a, b - y\} > 0.$$

Then $B_\varepsilon(x) = (x - \varepsilon, x + \varepsilon) \subset (a, b)$. Hence, (a, b) is a neighorhood of x for each $x \in (a, b)$, hence (a, b) is open in $\mathbb{R}$.

(2) Consider a closed interval $[a, b]$, where $a, b \in \mathbb{R}$ with $a < b$, For each $x \in (a, b)$, it is a neighborhood of x, which is shown as in (1), but it is not a neighborhood of a: For any $\varepsilon > 0$, $B_\varepsilon(a) = (a - \varepsilon, a + \varepsilon)$ contains $(a - \varepsilon, a)$, which is non-empty and disjoint from $[a, b]$. Thus, $[a, b]$ is *not* open in $\mathbb{R}$.

The same argument shows that half-open intervals $[a, +\infty)$ and $(-\infty, b]$ are *not* open in $\mathbb{R}$.

(3) Let (a_1, b_1) and (a_2, b_2) be two open intervals in $\mathbb{R}$. Then $U := (a_1, b_1) \times (a_2, b_2)$ is open in $\mathbb{R}^2$:

Let $(x_0, y_0) \in U$, i.e. so that $a_1 < x_0 < b_1$ and $a_2 < y_0 < b_2$. put

$$\varepsilon := \min\{x_0 - a_1, b_1 - x_0, y_0 - a_2, b_2 - y_0\} > 0.$$

Then for each $(x, y) \in B_\varepsilon(x_0, y_0)$ we compute that

$$|x - x_0| \leq \|(x, y) - (x_0, y_0)\| < \varepsilon \quad \text{and} \quad |y - y_0| < \varepsilon.$$

Hence, $x < x_0 + \varepsilon \leq x_0 + b_1 - x_0 = b_1$ and $x > x_0 - \varepsilon \geq x_0 - (x_0 - a_1) = a_1$, so $x \in (a_1, b_1)$. Analogously, one shows that $y \in (a_2, b_2)$, hence $(x, y) \in U$. Since $(x, y) \in B_\varepsilon(x_0, y_0)$ was chosen arbitrarily, $B_\varepsilon(x_0, y_0) \subset U$. Hence, U is a neighborhood of (x_0, y_0) and since (x_0, y_0) was chosen arbitrarily, U is open in $\mathbb{R}^2$.

Remark 11.48. The formalism of open balls in metric spaces allows for a concise description of continuity in a point. Namely, by definition, $f : X \to Y$ is continuous in a, where X and Y are metric spaces, if and only if

$$\forall \varepsilon > 0 \ \ \exists \delta > 0 : \quad f(B_\delta(a)) \subset B_\varepsilon(f(a)).$$

The following theorem collects some elementary set-theoretic properties of open subsets of metric spaces.

Theorem 11.49. *Let (X, d) be a metric space.*

a) $B_\varepsilon(x_0)$ is open in X for all $x_0 \in X$ and $\varepsilon > 0$.

b) $\varnothing$ and X are open in X.

c) Let I be a set. If $(U_i)_{i \in I}$ is a family of open subsets of X, then $\bigcup_{i \in I} U_i$ is open in X.

d) Let $n \in \mathbb{N}$. If $U_1, U_2, \ldots, U_n$ are open subsets of X, then $\bigcap_{i=1}^n U_i$ is open in X.

Proof. a) Let $x \in B_\varepsilon(x_0)$. We need to show that there exists $\delta > 0$, such that $B_\delta(x) \subset B_\varepsilon(x_0)$. Put $\delta := \varepsilon - d(x, x_0) > 0$. Then for all $a \in B_\delta(x)$ it holds by the triangle inequality that

$$d(a, x_0) \leq d(a, x) + d(x, x_0) < \delta + d(x, x_0) = \varepsilon,$$

hence $a \in B_\varepsilon(x_0)$. This shows that $B_\delta(x) \subset B_\varepsilon(x_0)$, which yields the claim.

b) For $\varnothing$ the condition on openness is empty, hence $\varnothing$ is open. For any $x \in X$ and $\varepsilon > 0$ it holds that $B_\varepsilon(x) \subset X$, hence X is open as well.

c) Let $x_0 \in \bigcup_{i \in I} U_i$. Then there exists $i_0 \in I$ with $x_0 \in U_{i_0}$ and since U_{i_0} is open, there exists $\varepsilon > 0$ with $B_\varepsilon(x_0) \subset U_{i_0} \subset \bigcup_{i \in I} U_i$. Hence, $\bigcup_{i \in I} U_i$ is open.

d) Let $x_0 \in \bigcap_{i=1}^n U_i$, i.e. $x_0 \in U_i$ for all $i \in \{1, 2, \ldots, n\}$. Since the U_i are open, for each $i \in \{1, 2, \ldots, n\}$ there exists $\varepsilon_i > 0$, such that $B_{\varepsilon_i}(x_0) \subset U_i$. Put $\varepsilon := \min\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\} > 0$. Then, since $\varepsilon_0 \leq \varepsilon_i$ for each i, it holds that $B_{\varepsilon_0}(x_0) \subset B_{\varepsilon_i}(x_0)$ and thus $B_{\varepsilon_0}(x_0) \subset U_i$ for each $i \in \{1, 2, \ldots, n\}$. Hence, $B_{\varepsilon_0}(x_0) \subset \bigcap_{i=1}^n U_i$. Since x_0 was chosen arbitrarily, it follows that $\bigcap_{i=1}^n U_i$ is open.

□

Remark 11.50. Infinite intersections of open sets are in general not open. While the interval $I_n = (-\frac{1}{n}, \frac{1}{n}) = B_{1/n}(0)$ is open in $\mathbb{R}$ for each $n \in \mathbb{N}$, we observe $\bigcap_{n \in \mathbb{N}} I_n = \bigcap_{n \in \mathbb{N}} (-\frac{1}{n}, \frac{1}{n}) = \{0\}$, which is not open in $\mathbb{R}$, as it contains no open ball of the form $B_r(0) = (-r, r)$.

Using the notion of open subsets, we can characterize the continuity of a map between metric spaces in a surprisingly concise way.

Theorem 11.51. *Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$. The map f is continuous if and only if $f^{-1}(U)$ is open in X for every $U \subset Y$ which is open in Y.*

Proof. "⇒": Let U be open in Y and let $x_0 \in f^{-1}(U)$. Since $f(x_0) \in U$ and U is open, there exists $\varepsilon > 0$ with $B_\varepsilon(f(x)) \subset U$. As carried out in Remark 11.48, since f is continuous, there is a $\delta > 0$ with $f(B_\delta(x_0)) \subset B(f(x_0)) \subset U$, hence $B \subset f^{-1}(f(B_\delta(x_0))) \subset f^{-1}(U)$, (where we used Theorem 1.17.a).(iii)). Thus, $f^{-1}(U)$ is a neighborhood of x_0 and since x_0 was chosen arbitrarily, this shows that $f^{-1}(U)$ is open in X.
"⇐": Let $x_0 \in X$ and let $\varepsilon > 0$. By Theorem 11.49.a), $B_\varepsilon(f(x_0))$ is open in Y, so by assumption $f^{-1}(B_\varepsilon(f(x_0))$ is open in X. Since $x_0 \in f^{-1}(f(x_0)) \subset f^{-1}(B_\varepsilon(f(x_0))$, it follows that there exists $\delta > 0$, such that $B_\delta(x_0) \subset f^{-1}(B_\varepsilon(f(x_0))$, which yields $f(B_\delta(x_0)) \subset B_\varepsilon(f(x_0))$. In other words, it holds for all $x \in X$ with $d_X(x, x_0) < \delta$ that $d_Y(f(x), f(x_0)) < \varepsilon$. As $\varepsilon > 0$ was arbitrary, this shows that f is continuous in x_0. □

We next introduce another property of sets whose use is not exactly evident upon first sight.

Definition 11.52. Let (X, d) be a metric space. $A \subset X$ is *closed in* X if $X \setminus A$ is open in X.

Our main purpose of discussing closed sets is in the following theorem, which characterizes closed subsets by convergent sequences.

Theorem 11.53. *Let (X, d) be a metric space and let $A \subset X$. Then A is closed in X if and only if for each sequence $(x_n)_{n\in\mathbb{N}}$ that converges in X it holds that $\lim\limits_{n\to\infty} x_n \in A$.*

Proof. "⇒": Assume that A is closed and that there was a sequence $(x_n)_{n\in\mathbb{N}}$ in A with $x_0 := \lim_{n\to\infty} x_n \notin A$. Since $X \setminus A$ is open and $x_0 \in X \setminus A$, there exists $\varepsilon > 0$ such that $B_\varepsilon(x_0) \subset X \setminus A$, i.e. $B_\varepsilon(x_0) \cap A = \varnothing$. In particular, $x_n \notin B_\varepsilon(x_0)$ for all $n \in \mathbb{N}$. But this is a contradiction since by definition of convergence, x_n must lie in $B_\varepsilon(x_0)$ for almost every n. Hence, $x_0 \in A$, which shows the claim.
"⇐": Let $x_0 \in X \setminus A$ and assume that $X \setminus A$ was not a neighborhood of x_0. Then for each $\varepsilon > 0$, it holds that $B_\varepsilon(x_0) \not\subset X \setminus A$, i.e. $B_\varepsilon(x_0) \cap A \neq \varnothing$. Thus, for each $n \in \mathbb{N}$ we can pick an $x_n \in B_{1/n}(x_0) \cap A$. Then one easily checks that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x_0. Since $(x_n)_{n\in\mathbb{N}}$ is a sequence in A, it holds by assumption that $x_0 = \lim_{n\to\infty} x_n \in A$, which contradicts the assumption that $x_0 \in X \setminus A$. Thus, $X \setminus A$ is a neighborhood of x_0 and since x_0 was chosen arbitrarily, it follows that $X \setminus A$ is open in X, i.e. that A is closed in X. □

Example 11.54. Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 < b_1$ and $a_2 < b_2$ and consider the rectangle $R = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$. Let $(x_n, y_n)_{n\in\mathbb{N}}$ be a sequence in R which converges in $\mathbb{R}^2$. By Theorem 11.7, both $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ converge in $\mathbb{R}$ and since $a_1 \leq x_n \leq b_1$ and $a_2 \leq y_n \leq b_2$ for all $n \in \mathbb{N}$ it holds that

$$a_1 \leq \lim_{n\to\infty} x_n \leq b_1 \quad \text{and} \quad a_2 \leq \lim_{n\to\infty} y_n \leq b_2.$$

Thus, $\lim\limits_{n\to\infty}(x_n, y_n) = (\lim\limits_{n\to\infty} x_n, \lim\limits_{n\to\infty} y_n) \in R$. As the sequence was chosen arbitrarily, it follows from Theorem 11.53 that R is closed in $\mathbb{R}^2$.

A general class of examples of closed subsets is discussed in part a) of the following theorem which collects set-theoretic properties. Note the similarity of Theorem 11.49 and Theorem 11.55. In fact, we will derive Theorem 11.55 as a consequence of Theorem 11.49 by studying complements of the constructions involved.

Theorem 11.55. *Let (X, d) be a metric space.*

a) The closed ball of radius ε around x_0, *i.e. the set*

$$\overline{B}_\varepsilon(x_0) := \{x \in X \mid d(x, x_0) \leq \varepsilon\}$$

is closed in X for all $x_0 \in X$ and $\varepsilon > 0$.

b) $\varnothing$ and X are closed in X.

c) Let I be a set. If $(A_i)_{i\in I}$ is a family of closed subsets of X, then $\bigcap_{i\in I} A_i$ is closed in X.

d) Let $n \in \mathbb{N}$. If $A_1, A_2, \ldots, A_n$ are closed subsets of X, then $\bigcup_{i=1}^{n} A_i$ is closed in X.

Proof. (This proof is not carried out in the lecture videos.)

a) We have to show that $U_\varepsilon := X \setminus B_\varepsilon(x_0) = \{x \in X \mid d(x, x_0) > \varepsilon\}$ is open in X. Let $x \in U_\varepsilon$ and put $\delta := d(x, x_0) - \varepsilon > 0$. For each $a \in B_\delta(x)$ we obtain from the triangle inequality that $d(x, x_0) \leq d(x, a) + d(a, x_0)$ and thus

$$d(a, x_0) \geq d(x, x_0) - d(x, a) > d(x, x_0) - \delta = \varepsilon.$$

Hence, $B_\delta(x) \subset U_\varepsilon$. Since x was chosen arbitrarily, it follows that U_ε is open in X.

b) This follows from Theorem 11.49.b) since $X \setminus \varnothing = X$ and $X \setminus X = \varnothing$ are open in X.

c) Since A_i is closed for each $i \in I$, it holds that $X \setminus A_i$ is open in X for each $i \in I$. A version of de Morgan's laws for unions of infinitely many sets, i.e. a generalization of Theorem 1.12.d) then yields that

$$X \setminus \bigcap_{i\in I} A_i = \bigcup_{i\in I} (X \setminus A_i),$$

which is open in X by Theorem 11.49.c). Hence $\bigcap_{i\in I} A_i$ is closed in X.

d) Since $A_1, \ldots, A_n$ are closed in X, the sets $X \setminus A_1, \ldots, X \setminus A_n$ are open in X. Iterating de Morgan's laws (Theorem 1.12.d)) n times then shows that

$$X \setminus \bigcup_{i=1}^{n} A_i = \bigcap_{i=1}^{n} (X \setminus A_i),$$

which is open in X by Theorem 11.49.d). Hence, $\bigcup_{i=1}^{n} A_i$ is closed in X.

□

We next discuss some additional constructions involving open and closed subsets that will come in handy in the next sections.

Definition 11.56. Let (X, d) be a metric space and let $A \subset X$.

a) A point $x_0 \in X$ is called a *boundary point of A* if for each $\varepsilon > 0$ it holds that

$$B_\varepsilon(x_0) \cap A \neq \varnothing \quad \text{and} \quad B_\varepsilon(x_0) \cap (X \setminus A) \neq \varnothing,$$

i.e. if every neighborhood of x_0 contains both an element of A and an element of $X \setminus A$. The *boundary of A* is given by

$$\partial A := \{x_0 \in X \mid x_0 \text{ is a boundary point of } A\}.$$

b) The *closure of A* is given by $\overline{A} := A \cup \partial A$, also denoted by $\mathrm{cl}(A)$.

c) The *interior of A* is given by $\mathring{A} := A \setminus \partial A$, also denoted by $\mathrm{int}(A)$.

Example 11.57. (1) Let $a, b \in \mathbb{R}$ with $a < b$ and let $I = [a, b) \subset \mathbb{R}$. Then

$$\mathring{I} = (a, b), \quad \overline{I} = [a, b] \quad \text{and} \quad \partial I = \{a, b\}.$$

Likewise, if $J = (a, b)$, then $\mathring{J} = J = (a, b)$, $\overline{J} = [a, b]$ and $\partial J = \{a, b\}$.

(2) Let $A \subset \mathbb{R}^2$ be given by

$$A = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\} \cup \{(x, y) \in \mathbb{R}^2 \mid x \geq 0,\ x^2 + y^2 = 1\}.$$

One checks that

$$\begin{aligned}
\mathring{A} &= \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}, && \text{(the interior of the unit circle)} \\
\overline{A} &= \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}, && \text{(the interior of the unit circle and the circular line)} \\
\partial A &= \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}. && \text{(the circular line of the unit circle)}
\end{aligned}$$

Remark 11.58. Let X be a metric space and let $A \subset X$.

(1) Note that by definition of interior and closure $\mathring{A} \subset A \subset \overline{A}$.

(2) The interior and the closure of A can be characterized by

$$\begin{aligned}
\mathring{A} &= \{x \in X \mid A \text{ is a neighborhood of } x\}, \\
\overline{A} &= \{x \in X \mid \exists (a_n)_{n \in \mathbb{N}} \subset A \text{ with } \lim_{n \to \infty} a_n = x\}.
\end{aligned}$$

These are straightforward consequences of the definitions and the reader is invited to carry out a formal and detailed proof of these statements.

The following theorem collects several basic results about boundaries, closures and interiors.

Theorem 11.59. *Let (X, d) be a metric space and let $A \subset X$. Then:*

a) $\mathring{A}$ is open in X.

b) ∂A and $\overline{A}$ are closed in X.

c) A is open in X if and only if $\mathring{A} = A$.

d) A is closed in X if and only if $\overline{A} = A$.

Proof. (This proof is not discussed in the lecture videos.)

a) Let $a \in \mathring{A} = A \setminus \partial A$. Then there exists $\varepsilon > 0$ such that $B_\varepsilon(a) \cap (X \setminus A) = \varnothing$, i.e. $B_\varepsilon(a) \subset A$. It remains to show that $B_\varepsilon(a) \subset \mathring{A}$, i.e. $B_\varepsilon(a) \cap \partial A = \varnothing$, as then $\mathring{A}$ is a neighborhood of a. Assume there was $x \in \partial A \cap B_\varepsilon(a)$. Since $B_\varepsilon(a)$ is open in X, there exists $\delta > 0$, such that $B_\delta(x) \subset B_\varepsilon(a) \subset A$. But this is a contradiction, since $x \in \partial A$. Hence, such an x does not exist, which shows the claim.

b) Let $A^c = X \setminus A$. It follows from the definition of a boundary that $\partial A = \partial A^c$. By a), $\operatorname{int}(A^c) = A^c \setminus \partial A^c = A^c \setminus \partial A$ is open, hence

$$X \setminus (A^c \setminus \partial A) = (X \setminus A^c) \cup \partial A = A \cup \partial A = \overline{A}$$

is closed in X. By some set-theoretic reasoning, one shows that

$$\partial A = \overline{A} \setminus \mathring{A} = \overline{A} \cap (X \setminus \mathring{A}).$$

Since $\overline{A}$ is closed and $X \setminus \mathring{A}$ is closed by a), it follows from Theorem 11.55.c) that ∂A is closed.

c) "⇒" is an immediate consequence of a).

"⇐": If A is open and $x \in A$, then there exists $\varepsilon > 0$ with $B_\varepsilon(x) \subset A$ and thus

$$B_\varepsilon(x) \cap (X \setminus A) = \varnothing.$$

Hence, $x \notin \partial A$. This shows that $A = A \setminus \partial A = \mathring{A}$.

d) "⇒" is an immediate consequence of b).

"⇐": Assume that $X \setminus A$ is open. Since $\partial A = \partial(X \setminus A)$, we derive from c) that

$$X \setminus A = \operatorname{int}(X \setminus A) = (X \setminus A) \setminus \partial(X \setminus A) = (X \setminus A) \setminus \partial A = X \setminus (A \cup \partial A) = X \setminus \overline{A}.$$

Taking complements yields $A = X \setminus (X \setminus A) = X \setminus (X \setminus \overline{A}) = \overline{A}$.

□

We next want to transfer the notions of continuous extensions of maps and limits of functions in a point from Chapter 3 of Mathematics 1 to the setting of metric spaces. The connection to closures of sets is made evident in Remark 11.61.(2) below.

Definition 11.60. Let X and Y be metric spaces and let $A \subset X$.

a) Let $x_0 \in \overline{A}$ and let $f : A \setminus \{x_0\} \to Y$ be continuous. We say that f *has a limit in* x_0 if there exists $y_0 \in Y$, such that

$$\tilde{f} : A \cup \{x_0\} \to Y, \qquad \tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in A \setminus \{x_0\}, \\ y_0 & \text{if } x = x_0, \end{cases}$$

In this case, we call y_0 *the limit of* f *in* x_0 and put $\lim\limits_{x \to x_0} f(x) := y_0$.

b) Let $f : A \to Y$ be continuous and let $B \subset X$ with $A \subset B \subset \overline{A}$. We say that $g : B \to Y$ is a *continuous extension of* f if g is continuous with $g|_A = f$.

Remark 11.61. (1) It follows from Theorem 11.30 about sequential continuity that

$$\lim_{x \to x_0} f(x) = b \quad \Leftrightarrow \quad \lim_{n \to \infty} f(x_n) = b \text{ for all sequences } (x_n)_{n \in \mathbb{N}} \text{ in } A \text{ with } \lim_{n \to \infty} x_n = a.$$

(2) Let $A \subset B \subset \overline{A}$. By definition, there is a a continuous extension of f to B if and only if f has a limit in every point in B.

To conclude this section, we will transfer the notion of uniform continuity from Chapter 3 to the setting of maps between metric spaces. We will then use this new notion to derive a useful result about continuous extensions of maps between metric spaces.

Definition 11.62. Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is *uniformly continuous* if

$$\forall \varepsilon > 0 \; \exists \delta > 0 : \;\; d_Y(f(x), f(y)) < \varepsilon \quad \forall x, y \in X \text{ with } d_X(x, y) < \delta.$$

Theorem 11.63. *Let (X, d_X) and (Y, d_Y) be metric spaces and let $A \subset X$. Assume that Y is complete. If $f : A \to Y$ is uniformly continuous, then f has a continuous extension to $\overline{A}$.*

Proof. (This proof is not discussed in the lecture videos.)

By Remark 11.61.(3), it suffices to show that f has a limit in every $a \in \overline{A}$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in A with $\lim_{n\to\infty} x_n = a$. We need to show that $(f(x_n))_{n\in\mathbb{N}}$ converges in Y and that its limit is independent of the choice of $(x_n)_{n\in\mathbb{N}}$. Let $\varepsilon > 0$. Since f is uniformly continuous, there exists $\delta > 0$, such that

$$d_X(x,y) < \delta \quad \Rightarrow d_Y(f(x), f(y)) < \varepsilon \qquad \forall x, y \in A.$$

Since $(x_n)_{n\in\mathbb{N}}$ converges to a, there exists $n_0 \in \mathbb{N}$, such that $d_X(x_n, a) < \frac{\delta}{2}$ for all $n \geq n_0$ and thus $d_X(x_n, x_m) \leq d_X(x_n, a) + d_X(x_m, a) < \frac{\delta}{2} + \frac{\delta}{2} = \delta$ for all $m, n \geq n_0$. Hence,

$$d_y(f(x_n), f(x_m)) < \varepsilon \qquad \forall m, n \geq n_0.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that $(f(x_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in Y and since Y is complete, there exists $b \in Y$ with $\lim_{n\to\infty} f(x_n) = b$. Let $(y_n)_{n\in\mathbb{N}}$ be another sequence in A converging to a. Analogously, we obtain that $\lim_{n\to\infty} f(y_n) = b'$ for some $b' \in Y$. Define a third sequence by

$$(z_n)_{n\in\mathbb{N}}, \quad z_n := \begin{cases} x_k & \text{if } n = 2k-1,\ k \in \mathbb{N}, \\ y_k & \text{if } n = 2k,\ k \in \mathbb{N}, \end{cases}$$

One computes that $\lim_{n\to\infty} z_n = a$ and as for the other two sequences we obtain that

$$\lim_{n\to\infty} f(z_n) = b''$$

for some $b'' \in Y$. Since both $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are subsequences of $(z_n)_{n\in\mathbb{N}}$ and since subsequences of convergent sequences converge to the same limit as the original sequence (this is shown along the same lines as Theorem 2.19), it follows that $b = b' = b''$. Thus, by Remark 11.61.(4), it follows that $\lim_{x\to a} f(x) = b$. This shows the claim. □

11.5 Compactness in metric spaces

We conclude our chapter on metric spaces by introducing the important notion of *compactness* of subsets of metric spaces. In Definition 3.26, we have called a subset of $\mathbb{R}$ or $\mathbb{C}$ compact if it is closed and bounded. The general definition of compact subsets of metric spaces is way more sophisticated and makes use of set-theoretic notions from the previous section. However, as we shall see in the Heine-Borel theorem, compact subsets of $\mathbb{R}^k$, which is the situation that we will primarily be interested in for the rest of this course, can be characterized in much simpler terms.

If you struggle at understanding the notion of compactness on this first encounter, please don't be discouraged by that. For the next chapters, it will (probably) be sufficient to understand the characterization of compact subsets of $\mathbb{R}^k$ by the Heine-Borel theorem and the statements of the theorems discussed in this section.

Definition 11.64. Let (X, d) be a metric space and let $A \subset X$.

a) A family $(U_i)_{i\in I}$, where I is a set, is called an *open cover of A* if U_i is open in X for every $i \in I$ and if

$$A \subset \bigcup_{i\in I} U_i.$$

b) A is *compact* if for *every* open cover $(U_i)_{i\in I}$ of A there exists a *finite* $J \subset I$, such that

$$A \subset \bigcup_{j\in J} U_j.$$

$(U_j)_{j\in J}$ is then called a *finite subcover of* $(U_i)_{i\in I}$.

While it is hard to derive explicit examples of compact sets straight from the definition, we can at least find a basic class of counterexamples before discussing a particular type of compact subsets in the subsequent theorem.

Example 11.65. Let V be a normed vector space with $V \neq \{0\}$. Then V is *not* compact in V: The family of open balls $(B_r(0))_{r\in(0,+\infty)}$ satisfies $V = \bigcup_{r\in(0,+\infty)} B_r(0)$, so it is an open cover of V. However, there is no finite subcover since if there were $r_1, \dots, r_m \in (0, +\infty)$ with $V \subset \bigcup_{i=1}^m B_{r_i}(0)$, then with $R := \max\{r_1, \dots, r_m\}$ it follows that $V \subset B_R(0)$. This is a contradiction, since for every $v \in V \setminus \{0\}$ it holds that $\frac{R}{\|v\|} v \notin B_R(0)$.

Theorem 11.66. *Let (X, d) be a metric space, let $(x_n)_{n\in\mathbb{N}}$ be a convergent sequence in X and put $a := \lim_{n\to\infty} x_n$. Then*

$$K := \{x_n \mid n \in \mathbb{N}\} \cup \{a\}$$

is a compact subset of K.

Proof. Let $(U_i)_{i\in I}$ be an open cover of K and let $i_0 \in I$ with $a \in U_{i_0}$. Since U_{i_0} is open, there exists $\varepsilon > 0$ with $B_\varepsilon(a) \subset U_{i_0}$. Since $(x_n)_{n\in\mathbb{N}}$ converges to A there exists $n_0 \in \mathbb{N}$ with $x_n \in B_\varepsilon(a)$ for all $n \geq n_0$, hence $\{x_n \mid n \geq n_0\} \cup \{a\} \subset U_{i_0}$. For each $j \in \{1, 2, \dots, n_0 - 1\}$ pick $i_j \in I$ with $x_j \in U_{i_j}$. Then by construction $K \subset \bigcup_{j=1}^{n_0-1} U_{i_j} \cup U_{i_0}$, so we have found a finite subcover of K. Since the open cover was chosen arbitrarily, this shows that K is compact. □

A crucial property of compactness it that it possesses an alternative characterization in terms of sequences. Both ways of defining compactness have their importance and their applications and we shall make use of both version in the course of this section.

Theorem 11.67. *Let (X, d) be a metric space and let $K \subset X$. The following statements are equivalent:*

(i) K is compact.

(ii) Every sequence in K has a subsequence which converges in K.

Before proving this result, we need to establish a little auxiliary lemma.

Lemma 11.68. *Let (X, d) be a metric space, let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X and let $a \in X$. If $B_\varepsilon(a)$ contains infinitely many elements of $(x_n)_{n\in\mathbb{N}}$ for all $\varepsilon > 0$, then $(x_n)_{n\in\mathbb{N}}$ has a subsequence that converges to a.*

Proof. We construct the desired subsequence inductively. Let $x_{n_1} \in B_1(a)$. Assuming that $x_{n_1}, \dots, x_{n_{k-1}}$ with $n_1 < n_2 < \cdots < n_{k-1}$ have been found so that $x_{n_j} \in B_{1/j}(a)$ for all $j \in \{1, 2, \dots, k-1\}$. Since $B_{1/k}(x_0)$ contains infinitely many elements of the sequence, there exists $n_k \in \mathbb{N}$ with $n_k > n_{k-1}$, so that $x_{n_k} \in B_{1/k}(x_0)$. The thus-constructed subsequence $(x_{n_k})_{k\in\mathbb{N}}$ then satisfies by definition that

$$\lim_{k\to\infty} d(x_{n_k}, a) \leq \lim_{k\to\infty} \frac{1}{k} = 0,$$

hence converges to a. □

Proof of Theorem 11.67. (i) $\Rightarrow$ (ii): Assume that there was a sequence $(x_n)_{n\in\mathbb{N}}$ in K which has no convergent subsequence. By Lemma 11.68, then for every $a \in K$ there is an $\varepsilon_a > 0$, such that $B_{\varepsilon_a}(a)$ contains only finitely many elements of $(x_n)_{n\in\mathbb{N}}$. Since $(B_{\varepsilon_a}(a))_{a\in K}$ is an open cover of K and K is compact, there are $a_1, \dots, a_m \in K$ with $K \subset \bigcup_{i=1}^m B_{\varepsilon_{a_i}}(a_i)$. But since each $B_{\varepsilon_{a_i}}(a_i)$ contains only finitely many of the x_n, so there can only be finitely many of the x_n in K, which is a contradiction as the whole sequence lies in K. Thus, $(x_n)_{n\in\mathbb{N}}$ has a subsequence that converges in K.

(ii) $\Rightarrow$ (i): *(This part of the proof is not discussed in the lecture videos.)*

The proof of this implication is based on the following statement:

Claim. Assume that (ii) holds. Then for all $\varepsilon > 0$ there are $m \in \mathbb{N}$ and $p_1, \dots, p_m \in K$ with $K \subset \bigcup_{i=1}^m B_\varepsilon(p_i)$.

Proof of the claim. Assume there was an $\varepsilon > 0$ for which the statement was false. Let $x_1 \in K$ be arbitrary. Then $K \not\subset B_\varepsilon(x_1)$, so pick $x_2 \in K \setminus B_\varepsilon(x_2)$. Proceedings recursively, we can find a sequence $(x_n)_{n\in\mathbb{N}}$, such that $x_n \in K \setminus \bigcup_{k=1}^{n-1} B_\varepsilon(x_k)$ Since $K \not\subset \bigcup_{k=1}^{n-1} B_\varepsilon(x_k)$ for any $n \in \mathbb{N}$, this sequence is well-defined and by construction satisfies

$$d(x_n, x_m) \geq \varepsilon \quad \forall m, n \in \mathbb{N} \text{ with } m \neq n.$$

This particularly shows that no subsequence of $(x_n)_{n\in\mathbb{N}}$ can be a Cauchy sequence. But since by assumption $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence, this contradicts Theorem 11.13. Hence, there is no such ε. ■

Assume that (ii) holds and let $(U_i)_{i\in I}$ be an arbitrary open cover of K which does not admit a finite subcover of K. We construct a sequence in the following way: Let $n \in \mathbb{N}$. By the previously proven claim there are $m_n \in \mathbb{N}$ and $x_{n,1}, x_{n,2}, \dots, x_{n,m_n} \in K$, such that $K \subset \bigcup_{j=1}^{m_n} B_{1/n}(x_{n,j})$. Since $(U_i)_{i\in I}$ has no finite subcover, there exists $j_0 \in \{1, 2, \dots, m_n\}$ such that $B_{1/n}(x_{n,j_0})$ is not covered by finitely many of the U_i. Put $x_n := x_{n,j_0}$. Performing this construction for all $n \in \mathbb{N}$, we obtain a sequence $(x_n)_{n\in\mathbb{N}}$. By (ii) the sequence has a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ which converges to some $a \in K$. Let $i_0 \in I$ such that $a \in U_{I_0}$. Since U_{i_0} is open, there exists $\varepsilon > 0$ with $B_\varepsilon(a) \subset U_{i_0}$. As $(x_n)_{n\in\mathbb{N}}$ converges to a, there is an $N \in \mathbb{N}$ with $\frac{1}{N} < \frac{\varepsilon}{2}$, such that $d(x_N, a) < \frac{\varepsilon}{2}$. But then for every $x \in B_{1/N}(x_n)$ it holds that

$$d(x, a) \leq d(x, x_N) + d(x_N, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus, $B_{1/N}(x_N) \subset B_\varepsilon(a) \subset U_{i_0}$, which is a contradiction to the definition of x_N. Thus, $(U_i)_{i\in I}$ must admit a finite subcover of K, which shows that K is compact. □

We will next use Theorem 11.67 to derive some basic properties of compact sets, among them the one introduced in the following definition.

Definition 11.69. Let (X, d) be a metric space. A subset $A \subset X$ is *bounded* if there exist $x_0 \in X$ and $R > 0$, such that $A \subset B_R(x_0)$.

Theorem 11.70. *Let (X, d) be a metric space and let $A \subset X$ be compact. Then A is closed and bounded.*

Proof. A is closed: Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in A which converges in X. By Theorem 11.67, $(x_n)_{n\in\mathbb{N}}$ then has a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ which converges to some $a \in A$. Since subsequences

of convergent sequences converge to the same limitm the sequence $(x_n)_{n\in\mathbb{N}}$ then itself converges to a, hence $\lim_{n\to\infty} x_n \in A$. Since the sequence was chosen arbitrarily, it follows from Theorem 11.53 that A is closed.
A is bounded: Fix $x_0 \in X$ and consider the family $(B_r(x_0))_{r\in(0,+\infty)}$. Every $B_r(x_0)$ is open in X and $X = \bigcup_{r\in(0,+\infty)} B_r(x_0)$. In particular, $(B_r(x_0))_{r\in(0,+\infty)}$ is an open cover of A and since A is compact there are $m \in \mathbb{N}$ and $r_1, r_2, \ldots, r_m \in (0, +\infty)$, such that $A \subset \bigcup_{j=1}^m B_{r_j}(x_0)$. But with $R := \max\{r_1, r_2, \ldots, r_m\}$ it holds that $\bigcup_{j=1}^m B_{r_j}(x_0) = B_R(x_0)$. This shows that A is bounded. □

For the elementary case of $X = \mathbb{R}$ the previous theorem has a direct consequence that allows us to characterize all compact intervals and to show that for intervals the definitions of compactness from Mathematics 1 (Definition 3.26) and Mathematics 2 (Definition 11.64) are compatible with one another.

Corollary 11.71. *Let $K \subset \mathbb{R}$ be compact (in the sense of Definition 11.64). Then* $\sup K \in K$ *and* $\inf K \in K$.

Proof. K is bounded by Theorem 11.70, hence $\sup K, \inf K \in \mathbb{R}$. K is closed by Theorem 11.70, and since there are sequences in K converging to $\sup K$ and to $\inf K$, it follows from Theorem 11.53 that $\sup K, \inf K \in K$. □

For subsets of $\mathbb{R}^k$, where $k \in \mathbb{N}$, compactness possesses a simple characterization which neither requires the study of open covers nor the study of convergent subsequences of arbitrary sequences. To establish such a characterization, we first generalize an old friend of ours from $\mathbb{R}$ to $\mathbb{R}^k$ for arbitrary $k \in \mathbb{N}$.

Theorem 11.72 (Bolzano-Weierstraß theorem for $\mathbb{R}^k$)**.** *Let $k \in \mathbb{N}$ and consider $\mathbb{R}^k$ with the Euclidean metric. Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. We prove the claim by induction over k. The base case of $k = 1$ is the one-dimensional Bolzano-Weierstraß theorem from Mathematics 1 (Theorem 2.48).
For the induction step, we assume as induction hypothesis that every bounded sequence in $\mathbb{R}^k$ converges for some $k \in \mathbb{N}$. Let $(x_{1n}, \ldots, x_{kn}, x_{(k+1)n})_{n\in\mathbb{N}}$ be a bounded sequence in $\mathbb{R}^{k+1}$ and let $R \geq 0$, such that $\|(x_{1n}, \ldots, x_{kn}, x_{(k+1)_n})\| \leq R$ for all $n \in \mathbb{N}$. Then

$$\|(x_{1n}, \ldots, x_{kn})\|^2 = \sum_{i=1}^{k} x_{in}^2 \leq \sum_{i=1}^{k+1} x_i^2 = \|(x_{1n}, \ldots, x_{kn}, x_{(k+1)n})\|^2 \leq R^2 \quad \forall n \in \mathbb{N}.$$

Hence $(x_{1n}, \ldots, x_{kn})_{n\in\mathbb{N}}$ is a bounded sequence in $\mathbb{R}^k$, thus has a convergent subsequence $(x_{1n_\ell}, \ldots, x_{kn_\ell})_{\ell\in\mathbb{N}}$ by the induction hypothesis. Since $|x_{(k+1)n}| \leq \|(x_{1n}, \ldots, x_{kn}, x_{(k+1)n})\| \leq R$ for all $n \in \mathbb{N}$, the sequence $(x_{(k+1)n})_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}$ and so is its subsequence $(x_{(k+1)n_\ell})_{\ell\in\mathbb{N}}$. Thus, by Theorem 2.48 it has a convergent subsequence $(x_{(k+1)n_{\ell_m}})_{m\in\mathbb{N}}$. Since $(x_{1n_\ell}, \ldots, x_{kn_\ell})_{\ell\in\mathbb{N}}$ is a subsequence of a convergent sequence and thus itself convergent, it follows from Theorem 11.7 that $(x_{1n_{\ell_m}}, \ldots, x_{kn_{\ell_m}}, x_{(k+1)n_{\ell_m}})_{m\in\mathbb{N}}$ converges in $\mathbb{R}^{k+1}$. This completes the induction step. □

Theorem 11.73 (Heine-Borel)**.** *Let $k \in \mathbb{N}$ and let $K \subset \mathbb{R}^k$. Then K is compact if and only if K is closed and bounded.*

Proof. Since "⇒" is a special case of Theorem 11.70, it only remains to show the converse implication.

Let $K \subset \mathbb{R}^k$ be closed and bounded. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in K. Since K is bounded, the sequence is bounded, so by the Bolzano-Weierstraß theorem (Theorem 2.48) it has a convergent subsequence $(x_{n_\ell})_{\ell\in\mathbb{N}}$. Since K is closed, it follows from Theorem 11.53 that $\lim_{\ell\to\infty} x_{n_\ell} \in K$. Thus, $(x_n)_{n\in\mathbb{N}}$ has a subsequence converging in K. Since $(x_n)_{n\in\mathbb{N}}$ was chosen arbitrarily in K, it follows from Theorem 11.67 that K is compact. □

Example 11.74. (1) Every closed ball $\overline{B}_r(x_0)$ in $\mathbb{R}^k$, where $r > 0$ and $x_0 \in \mathbb{R}^k$, is compact.

(2) An interval $I \subset R$ is compact if and only if $I = [a, b]$ for some $a, b \in \mathbb{R}$.

(3) More generally, every product of compact intervals in $\mathbb{R}^k$, i.e. every set of the form

$$Q = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_k, b_k] \subset \mathbb{R}^k$$

is compact, where $a_1, b_1, \ldots, a_k, b_k \in \mathbb{R}$.

Remark 11.75. For general metric spaces the statement of the Heine-Borel is not necessarily true, i.e. there are closed and bounded subsets of metric spaces which are not compact. Consider $(\mathcal{B}(\mathbb{R}, \mathbb{R}), \|\cdot\|_\infty)$ and put $A := \overline{B}_1(0)$, where 0 denotes the zero function. Explicitly,

$$A := \{f \in \mathcal{B}(\mathbb{R}, \mathbb{R}) \mid \sup_{x\in\mathbb{R}} |f(x)| \leq 1\}.$$

A is bounded by definition and closed by Theorem 11.55.a), but it is not compact: Consider the sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{B}(\mathbb{R}, \mathbb{R})$, where

$$f_n(x) = \begin{cases} 1 & \text{if } x = n, \\ 0 & \text{if } x \neq n. \end{cases}$$

Then $\|f_n\|_\infty = 1$ for all $n \in \mathbb{N}$, hence $(f_n)_{n\in\mathbb{N}}$ is a sequence in A. But since $\|f_n - f_m\|_\infty = 1$ whenever $n \neq m$, no subsequence of $(f_n)_{n\in\mathbb{N}}$ can be a Cauchy sequence. By Theorem 11.13, this implies that $(f_n)_{n\in\mathbb{N}}$ has no convergent subsequence, so by Theorem 11.67 A is not compact.

Theorem 11.76. *Let (X, d_X) and (Y, d_Y) be metric spaces. If $K \subset X$ is compact and $f : X \to Y$ is continuous, then $f(K) \subset Y$ is compact.*

Proof. Let $(V_i)_{i\in I}$ be an open cover of $f(K)$ in Y. Since f is continuous, it follows from Theorem 11.51 that $(f^{-1}(V_i))_{i\in I}$ is an open cover of K. Since K is compact, there are $m \in \mathbb{N}$, $i_1, i_2, \ldots, i_m \in I$ such that $K \subset \bigcup_{j=1}^m f^{-1}(V_{i_j})$, hence $f(K) \subset \bigcup_{j=1}^m f(f^{-1}(V_{i_j})) = \bigcup_{j=1}^m V_{i_j}$. Hence, $(V_i)_{i\in I}$ admits a finite subcover of $f(K)$. Thus, $f(K)$ is compact. □

Corollary 11.77. *Let (X, d) be a compact metric space and let $f : X \to \mathbb{R}$ be continuous. then f is bounded and attains its maximum and minimum, i.e. there are $x_m, x_M \in X$ with*

$$f(x_M) = \sup\{f(x) \mid x \in X\} \quad \text{and} \quad f(x_m) = \inf\{f(x) \mid x \in X\}.$$

Proof. By Theorem 11.76, $f(X)$ is compact in $\mathbb{R}$, hence $\sup f(X) \in f(X)$ and $\inf f(X) \in X$ by Corollary 11.71. This is equivalent to the claim. □

We conclude this section by presenting a theorem which is a direct generalization of Theorem 3.38 from Mathematics 1. Its proof will be part of an exercise sheet, thus be omitted here.

Theorem 11.78. *Let (X, d_X) and (Y, d_Y) be metric spaces. If X is compact and $f : X \to Y$ is continuous, then f is uniformly continuous.*

Proof. Sheet 11, Exercise 4. □

Chapter 12

Differentiation in several real variables

12.1 Directional and partial derivatives

After establishing a very general notion of continuity in the framework of metric spaces in the previous chapter, we are going to focus on maps of the form $U \to \mathbb{R}^m$ with $U \subset \mathbb{R}^n$ for arbitrary $m, n \in \mathbb{N}$. More precisely, we will introduce a notion of differentiability for such maps in this chapter. Many of the constructions can be generalized to or have analogues for arbitrary normed vector spaces, but we will refrain from giving the most general constructions of the definitions in this chapter. We further use the following conventions throughout this chapter.

- We always consider $\mathbb{R}^n$ as equipped with the Euclidean norm and its induced metric.
- $(e_1, \dots, e_n)$ always denotes the canonical basis of $\mathbb{R}^n$.
- A set $U \subset \mathbb{R}^n$ which is open in $\mathbb{R}^n$ will simply be called *open*. We use the analogous convention for closed, bounded and compact subsets of $\mathbb{R}^n$.

In this section we want to take the first steps towards differentiation of maps between $\mathbb{R}^n$ and $\mathbb{R}^m$ by introducing directional and partial derivatives. While these notions do not yet give the full picture of differentiation, they are straightforward generalizations of the derivative of a real function, so we can use our knowledge from Mathematics 1 to make some first observations and computations for maps between subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$.

We recall from Definition 4.1 that a function $f : I \to \mathbb{R}$ is differentiable in $a \in I$, where I is an interval, if $f'(a) = \lim_{h \to 0} \frac{f(a+h)-f(a)}{h}$ exists in $\mathbb{R}$. Thinking about this construction, one realizes that this notion can be generalized in a straightforward way to maps whose domains are intervals, but which map to $\mathbb{R}^m$ for an arbitrary $m \in \mathbb{N}$.

Definition 12.1. Let I be an interval, $m \in \mathbb{N}$ and $f : I \to \mathbb{R}^m$.

a) If f is continuous, then f is called a *curve in* $\mathbb{R}^m$ or *path in* $\mathbb{R}^m$.

b) f is *differentiable in* $x_0 \in I$ if the limit

$$f'(x_0) := \lim_{h \to 0} \frac{1}{h}(f(x_0 + h) - f(x_0))$$

exists. Then $f'(x_0) \in \mathbb{R}^m$ is called the *derivative* or *tangent vector of f in x_0*.

c) f is *differentiable* if f is differentiable in every $x_0 \in I$.

d) f is *continuously differentiable* or *of class* C^1 if f is differentiable and if the map $f' : I \to \mathbb{R}^m$ is continuous.

Remark 12.2. As you have already seen in Theoretical Physics 1, curves in $\mathbb{R}^n$ are used in Newtonian mechanics to describe motions of particles or objects. The variable is interpreted as a time parameter and for a differentiable curve $\gamma : I \to \mathbb{R}^n$, the point $\gamma(t_0)$ is the position of the particle/object at time t_0, while $\gamma'(t_0)$ describes its velocity at time t_0.

The question remains of how to explicitly compute derivatives of differentiable curves. The following theorem shows that the derivative can be taken componentwise, i.e. that we view the component functions of a curve $f : I \to \mathbb{R}^n$ as maps $I \to \mathbb{R}$ and can compute the derivatives of these components individually. As these are functions mapping to $\mathbb{R}$, we can then use all computational methods and rules for differentiation that we learnt about in Chapter 4 in Mathematics 1.

Theorem 12.3. *Let I be an interval, $m \in \mathbb{N}$ and let $f : I \to \mathbb{R}^m$, $f = (f_1, \dots, f_m)$, where $f_i : I \to \mathbb{R}$ for all $i \in \{1, 2, \dots, m\}$. Then f is differentiable in $a \in I$ if and only if f_i is differentiable in a for all $i \in \{1, 2, \dots, m\}$. Moreover, in case of differentiability,*

$$f'(a) = (f_1'(a), \dots, f_m'(a)).$$

Proof. We compute for all $h \in \mathbb{R}$ with $h \neq 0$ and $a + h \in I$ that

$$\begin{aligned}\frac{1}{h}(f(a+h) - f(a)) &= \frac{1}{h}((f_1(a+h), \dots, f_m(a+h)) - (f_1(a), \dots, f_m(a))) \\ &= \left(\frac{f_1(a+h) - f_1(a)}{h}, \dots, \frac{f_m(a+h) - f_m(a)}{h}\right).\end{aligned}$$

From this equation and Theorem 11.41 we derive that $\frac{1}{h}(f(a+h) - f(a))$ as a map in the variable h has the limit $b = (b_1, \dots, b_m)$ in 0 if and only if $\frac{f_i(a+h) - f_i(a)}{h}$ has the limit b_i in 0 for each $i \in \{1, 2, \dots, m\}$. By definition of differentiability, this shows the claim. □

Example 12.4. Consider the curve $\gamma : \mathbb{R} \to \mathbb{R}^3$, $\gamma(t) = (\cos(2\pi t), \sin(2\pi t), t)$. The graph of γ is a so-called *helical line* in $\mathbb{R}^3$. Viewing γ as describing a motion in dependence of time, in the first two coordinates, the curve is performing a periodic motion around the circle of radius 1 at constant speed. One sees that $\gamma(k) = (1, 0, k)$ for all $k \in \mathbb{Z}$ and that the third coordinate is just increasing at constant speed with t. By Theorem 12.3, γ is differentiable with

$$\gamma'(t) = (-2\pi \sin(2\pi t), 2\pi \cos(2\pi t), 1) \qquad \forall t \in \mathbb{R}.$$

The definition of derivatives of curves has no obvious generalization to maps of the form $f : U \to \mathbb{R}^m$, where $U \subset \mathbb{R}^n$ and $n > 1$. If we would view the h in $f(x_0 + h)$ occuring in Definition 12.1 as an element of $\mathbb{R}^n$, then $\frac{1}{h}$ was not well-defined, so we can not define any derivative by the same expression as in Definition 12.1.
A first idea at defining derivatives of such maps is to study only "parts of f", namely restrictions of f to lines going through the point under consideration. Given a point $a \in U$, any line in $\mathbb{R}^n$ going through a is of the form $\{a + tv \in \mathbb{R}^n \mid t \in \mathbb{R}\}$, where $v \in \mathbb{R}^n \setminus \{0\}$. So by restricting f to such a line and viewing the line as parametrized by one real variable t, we can interpret the restriction of f as a curve and study its derivative. This idea is formalized in the following definition.

Definition 12.5. Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be a map. Let $a \in U$, $v \in \mathbb{R}^n \setminus \{0\}$ and let $\varepsilon > 0$ be so small[1] that $a + tv \in U$ for all $t \in (-\varepsilon, \varepsilon)$. If the curve

$$g_v : (-\varepsilon, \varepsilon) \to \mathbb{R}^m, \qquad g_v(t) = f(a + tv),$$

is differentiable in 0, then we put

$$\partial_v f(a) := g_v'(0) = \lim_{t \to 0} \frac{1}{t}(f(a + tv) - f(a))$$

and call $\partial_v f(a) \in \mathbb{R}^m$ the directional derivative of f in a in direction v.

Example 12.6. (1) We consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1^2 + x_2^2$. We want to check the existence of $\partial_v f(a)$ for given $a = (a_1, a_2), v = (v_1, v_2) \in \mathbb{R}^2$. We compute for $t \in \mathbb{R}$ that

$$\begin{aligned} f(a + tv) - f(a) &= f(a_1 + tv_1, a_2 + tv_2) - f(a_1, a_2) = (a_1 + tv_1)^2 + (a_2 + tv_2)^2 - a_1^2 - a_2^2 \\ &= a_1^2 + 2ta_1v_1 + t^2v_1^2 + a_2^2 + 2ta_2v_2 + t^2v_2^2 - a_1^2 - a_2^2 \\ &= t(2a_1v_1 + 2a_2v_2) + t^2(v_1^2 + v_2^2). \end{aligned}$$

Consequently,

$$\partial_v f(a) = \lim_{t \to 0} \frac{1}{t}(f(a + tv) - f(a)) = \lim_{t \to 0}(2a_1v_1 + 2a_2v_2 + t(v_1^2 + v_2^2)) = 2a_1v_1 + 2a_2v_2.$$

Hence, all directional derivatives exist with $\partial_v f(a) = 2a_1v_1 + 2a_2v_2$ for all $v = (v_1, v_2) \in \mathbb{R}^2$.

(2) Consider $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x, y) \neq (0, 0), \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$

Let $a = (0, 0)$ and consider the vectors $v = (1, 0)$ and $w = (1, 1)$. One observes that

$$f((0,0) + tv) - f(0,0) = f(t, 0) = 0, \quad f((0,0) + tw) - f(0,0) = f(t, t) = \frac{t^2}{t^2 + t^2} = \frac{1}{2}$$

for all $t \in \mathbb{R}$. One derives from these computations that $\partial_v f(0, 0) = 0$, while $\partial_w f(0, 0)$ does not exist.

Directional derivatives with direction an element of the canonical basis of $\mathbb{R}^n$ are particularly interesting as we shall see in the course of this chapter. We introduce some additional terminology for these derivatives.

Definition 12.7. Let $m, n \in \mathbb{N}$ and let $U \subset \mathbb{R}^n$ be open. Let $f : U \to \mathbb{R}^m$ be a map and let $a = (a_1, \ldots, a_n) \in U$.

a) Let $i \in \{1, 2, \ldots, n\}$. If the directional derivative $\partial_{e_i} f(a)$ exists, then we put

$$\partial_i f(a) := \partial_{e_i} f(a)$$

and call $\partial_i f(a)$ *the i-th partial derivative of f in a*. Other notations for $\partial_i f(a)$ that occur in the literature include $\partial_{x_i} f(a)$, $D_i f(a)$, $f_{x_i}(a)$, $\frac{\partial f}{\partial x_i}(a)$, $\frac{\partial}{\partial x_i} f(a)$, ...

[1] Such an ε exists since U is open.

b) We call f *partially differentiable in* a if $\partial_i f(a)$ exists for every $i \in \{1, 2, \dots, n\}$.

c) We call f *partially differentiable* if f is partially differentiable in every $a \in U$.

Remark 12.8. By definition of the canonical basis, we compute in the notation of Definition 12.7 that

$$\partial_i f(a) = \lim_{t \to 0} \frac{1}{t}(f(a + te_i) - f(a)) = \lim_{t \to 0} \frac{1}{t}(f(a_1, \dots, a_{i-1}, a_i + t, a_{i+1}, \dots, a_n) - f(a_1, \dots, a_n)).$$

So with $b := (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n) \in \mathbb{R}^{n-1}$ fixed and with

$$g_b : (x - \delta, x + \delta) \to \mathbb{R}^m, \quad g_b(x) := f(a_1, \dots, a_{i-1}, x, a_{i+1}, \dots, a_n),$$

where $\delta > 0$ is choisen suitably small, it holds that

$$\partial_i f(a) = \lim_{t \to 0} \frac{1}{t}(g_b(a_i + t) - g_b(a_i)) = g_b'(a_i),$$

in the sense of Definition 12.1. Thus, we can compute partial derivatives by computing derivatives of curves, i.e. by using all of our one-dimensional rules of differentiation in each components to compute partial derivatives. Informally speaking, we just need to "derive the component functions of f by x_i" to compute the i-th partial derivative of f.

Example 12.9. (1) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function from Example 12.6.(2). In that example we have shown that $\partial_1 f(0,0) = 0$ and completely analogously, one shows that $\partial_2 f(0,0) = 0$. Hence, f is partially differentiable in $(0,0)$.

(2) Consider the map

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \qquad f(x_1, x_2) = \left(e^{-x_1^2 - x_2^2}, \sin(x_1 x_2)\right).$$

By our knowledge from Mathematics 1 and Theorem 12.3, for fixed $x_2 \in \mathbb{R}$ the curve $f_{x_2} : \mathbb{R} \to \mathbb{R}$, $f_{x_2}(x_1) = (e^{-x_1^2 - x_2^2}, \sin(x_1 x_2))$, is differentiable. Thus, by the chain rule,

$$\partial_1 f(x_1, x_2) = f_{x_2}'(x_1) = \left(-2x_1 e^{-x_1^2 - x_2^2}, x_2 \cos(x_1 x_2)\right) \qquad \forall (x_1, x_2) \in \mathbb{R}^2,$$

and analogously

$$\partial_2 f(x_1, x_2) = \left(-2x_2 e^{-x_1^2 - x_2^2}, x_1 \cos(x_1 x_2)\right) \qquad \forall (x_1, x_2) \in \mathbb{R}^2.$$

Hence, f is partially differentiable.

As discussed above, taking directional or partial derivatives of a map f amounts to only considering a part of the map at a time, namely a restriction to a line. As seen in the examples, some directional derivatives of a function might exist, while others might not. This suggests that by just studying selected directional derivatives, we just get "isolated" impressions of a map in certain directions, but do not grasp features of the map as a whole. Thus, to introduce a proper generalization of derivatives to higher dimensions, we still need to find a more elaborate construction for which we will use notions from linear algebra that we learnt about earlier in this course.

12.2 The differential of a map

In the previous semester, we have often emphasized the viewpoint of seeing ***differentiation as linear approximation*** of a map. More precisely, if I is an open interval and $f : I \to \mathbb{R}$ is differentiable in $a \in I$, then we have observed in Theorem 4.6 that f is of the form

$$f(a+h) = f(a) + f'(a) \cdot h + \varphi(h), \qquad \text{where } \lim_{h \to 0} \frac{\varphi(h)}{|h|} = 0.$$

This can be rewritten as $f(a+h) = f(a) + A(h) + \varphi(h)$, where $A : \mathbb{R} \to \mathbb{R}$, $A(h) = f'(a) \cdot h$. Meanwhile, we have learnt a great deal about linear algebra and as A is multiplication by a constant, A is a *linear map* from $\mathbb{R}$ to $\mathbb{R}$, where we view $\mathbb{R}$ as a vector space over itself. As a consequence of Section 8.1, more precisely of Theorem 8.5.b) for $n = m = 1$, we observe that *all* linear maps $\mathbb{R} \to \mathbb{R}$ are given by multiplication with a constant. Thus, $f'(a)$ is the real number for which ***multiplication with $f'(a)$ is the best approximation of $f : I \to \mathbb{R}$ by a linear map near a.*** With this property in mind, we next introduce the differential of a map as a generalization of that idea: as the best approximation of a map near a point by a linear map.

Definition 12.10. Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$.

a) Let $a \in U$. f is *differentiable in a* if there exists a *linear* map $A : \mathbb{R}^n \to \mathbb{R}^m$, such that

$$f(a+h) = f(a) + A(h) + \varphi(h) \qquad \forall h \in U',$$

where $U' := \{h \in \mathbb{R}^n \mid a + h \in U\}$ and where $\varphi : U' \to \mathbb{R}^m$ satisfies

$$\lim_{h \to 0} \frac{1}{\|h\|} \varphi(h) = 0.$$

We then put $Df_a := A \in L(\mathbb{R}^n, \mathbb{R}^m)$ and call Df_a *the differential of f in a*. Several different notations are used throughout the literature, the differential of f in a is also denoted e.g. by df_a, $T_a f$ or by $f'(a)$.

b) We call f *differentiable* if f is differentiable in every $a \in U$. In this case, we put

$$Df : U \to L(\mathbb{R}^n, \mathbb{R}^m), \qquad Df(a) := Df_a,$$

and call Df the *differential of f*.

Remark 12.11. The differential of a map $f : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ open, in $a \in U$ is well-defined, i.e. if there exists a linear map having the defining property, then it is unique: *(The next argument is not discussed in the lecture videos.)*
Let $A_1, A_2 \in L(\mathbb{R}^n, \mathbb{R}^m)$ and $\varphi_1, \varphi_2 : U' \to \mathbb{R}^m$, where $U' := \{h \in \mathbb{R}^n \mid a + h \in U\}$, such that

$$f(a+h) = f(a) + A_i(h) + \varphi_i(h) \qquad \forall h \in U',\ i \in \{1,2\}$$

where $\lim_{h \to 0} \frac{1}{\|h\|} \varphi_i(h) = 0$ for $i \in \{1,2\}$. Taking the difference of the two equations shows that

$$A_1(h) - A_2(h) = \varphi_2(h) - \varphi(h) \qquad \forall h \in U'.$$

Let $v \in \mathbb{R}^n$ be arbitrary and let $t > 0$ be so small that $t \cdot v \in U'$. Then by linearity of A_1 and A_2

$$\begin{aligned} & t(A_1(v) - A_2(v)) = A_1(tv) - A_2(tv) = \varphi_2(tv) - \varphi_1(tv) \\ \Leftrightarrow \quad & A_1(v) - A_2(v) = \frac{1}{t}(\varphi_2(tv) - \varphi_1(tv)) \end{aligned}$$

for all sufficiently small $t > 0$. Hence,

$$\begin{aligned} A_1(v) - A_2(v) &= \lim_{t \searrow 0} \frac{1}{t}(\varphi_2(tv) - \varphi_1(tv)) = \|v\| \lim_{t \searrow 0} \frac{1}{\|tv\|}(\varphi_2(tv) - \varphi_1(tv)) \\ &= \|v\| \left(\lim_{t \searrow 0} \frac{1}{\|tv\|} \varphi_2(tv) - \lim_{t \searrow 0} \frac{1}{\|tv\|} \varphi_1(tv) \right) = 0. \end{aligned}$$

Since $v \in \mathbb{R}^n \setminus \{0\}$ was chosen arbitrarily, this shows that $A_1 = A_2$.

As in the one-dimensional case, differentiability at a point implies continuity at a point.

Theorem 12.12. *Let $m, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^m$ and $a \in U$. If f is differentiable in a, then f is continuous in a.*

Proof. We will use Theorem 11.30 about sequential continuity. Let $(x_n)_{n \in NN}$ be a sequence in U with $\lim_{n \to \infty} x_n = a$. Put $h_n := x_n - a$ for all $n \in \mathbb{N}$, such that $\lim_{n \to \infty} h_n = 0$. By Corollary 11.42, Df_a is continuous, such that $\lim_{n \to \infty} Df_a(h_n) = Df_a(0) = 0$. Moreover, if φ is given as in Definition 12.10, $\lim_{n \to \infty} \varphi(h_n) = 0$ by construction. Hence,

$$\lim_{n \to \infty} f(x_n) = \lim_{n \to \infty} f(a + h_n) = \lim_{n \to \infty} (f(a) + Df_a(h_n) + \varphi(h_n)) = f(a).$$

Since $(x_n)_{n \in \mathbb{N}}$ was chosen arbitrarily, it follows from Theorem 11.30 that f is continuous in a. □

Remark 12.13. The analogue of Theorem 12.12 for *partially* differentiable function is false in general. The function f from Example 12.6.(2) is partially differentiable in $(0, 0)$, but not continuous in $(0, 0)$ as we have already seen in Example 11.31.(2).

Theorem 12.14. *Let $m, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$, $f = (f_1, \ldots, f_m)$, where $f_i : U \to \mathbb{R}$ for all $i \in \{1, 2, \ldots, m\}$. Then f is differentiable in $a \in U$ if and only if f_i is differentiable in a for each $i \in \{1, 2, \ldots, m\}$. In that case, it further holds that*

$$Df_a(v) = ((Df_1)_a(v), \ldots, (Df_m)_a(v)) \qquad \forall v \in \mathbb{R}^n.$$

Proof. (This proof is not discussed in the lecture videos.)
"⇒": Let f be differentiable in a. For $i \in \{1, 2, \ldots, m\}$ we define $A_i : \mathbb{R}^n \to \mathbb{R}$ by demanding that $Df_a(v) = (A_1(v), \ldots, A_m(v))$ for all $v \in \mathbb{R}^n$. By linearity of Df_a, each A_i is linear. We need to show that f_i is differentiable in a with differential A_i for each i. One checks from the definition of differentiability that this is the case if and only if

$$\lim_{h \to 0} \frac{|f_i(a + h) - f_i(a) - A_i(h)|}{\|h\|} \stackrel{!}{=} 0 \quad \forall i \in \{1, 2, \ldots, m\}.$$

Since by definition

$$f(a + h) - f(a) - Df_a(h) = (f_1(a + h) - f_1(a) - A_1(h), \ldots, f_m(a + h) - f_m(a) - A_m(h))$$

for all suitable $h \in \mathbb{R}^n$, it follows from a standard estimate that

$$|f_i(a + h) - f_i(a) - A_i(h)| \leq \|f(a + h) - f(a) - Df_a(h)\|$$

for all i and all suitable $h \in \mathbb{R}^n$. Hence,

$$\lim_{h\to 0} \frac{|f_i(a+h) - f_i(a) - A_i(h)|}{\|h\|} \leq \lim_{h\to 0} \frac{\|f(a+h) - f(a) - Df_a(h)\|}{\|h\|}$$
$$= \lim_{h\to 0} \left\| \frac{1}{\|h\|}(f(a+h) - f(a) - Df_a(h)) \right\| = 0.$$

and thus $\lim_{h\to 0} \frac{|f_i(a+h)-f_i(a)-A_i(h)|}{\|h\|} = 0$, which in turn yields $\lim_{h\to 0} \frac{1}{\|h\|}(f_i(a+h) - f_i(a) - A_i(h)) = 0$ for all $i \in \{1, 2, \ldots, m\}$ by the squeeze theorem. Hence, every f_i is differentiable in a with $(Df_i)_a = A_i$.
"⇐": Assume that f_i is differentiable in a for each $i \in \{1, \ldots, m\}$ and define $A : \mathbb{R}^n \to \mathbb{R}^m$ by $A(v) = ((Df_1)_a(v), \ldots, (Df_m)_a(v))$. Since the $(Df_i)_a$ are linear, A is linear. We compute from the differentiability assumption that

$$\lim_{h\to 0} \left\| \frac{1}{\|h\|}(f(a+h) - f(a) - A(h)) \right\| = \lim_{h\to 0} \frac{\|f(a+h) - f(a) - A(h)\|}{\|h\|}$$
$$= \lim_{h\to 0} \frac{\sqrt{\sum_{i=1}^n (f_i(a+h) - f_i(a) - Df_i(a))^2}}{\|h\|} \leq \lim_{h\to 0} \frac{\sum_{i=1}^n |f_i(a+h) - f_i(a) - Df_i(a)|}{\|h\|}$$
$$= \sum_{i=1}^n \lim_{h\to 0} \frac{|f_i(a+h) - f_i(a) - Df_i(a)|}{\|h\|} = \sum_{i=1}^n \lim_{h\to 0} \left| \frac{1}{\|h\|}(f_i(a+h) - f_i(a) - Df_i(a)) \right| = 0,$$

which yields $\lim_{h\to 0} \frac{1}{\|h\|}(f(a+h) - f(a) - A(h)) = 0$. Thus, f is differentiable in a with $Df_a = A$. □

The obvious question is now how the differential of a map in a point is related to directional and partial derivatives that we constructed independently. The following important theorem establishes a relation between the two. It turns out that *the differential of a differentiable map in a point contains the information about all directional derivatives in that point.*

Theorem 12.15. *Let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be differentiable in $a \in U$. Then*

$$Df_a(v) = \partial_v f(a) \qquad \forall v \in \mathbb{R}^n \setminus \{0\}.$$

In particular, all directional derivatives of f in a exist.

Proof. Let $v \in \mathbb{R}^n \setminus \{0\}$ and put $g_v : (-\varepsilon, \varepsilon) \to \mathbb{R}^m$, $g_v(t) = f(a + tv)$, for suitable $\varepsilon > 0$. Then for all $t \neq 0$:

$$\frac{1}{t}(g_v(t) - g_v(0)) = \frac{1}{t}(f(a+tv) - f(a)) = \frac{1}{t}(Df_a(tv) + \varphi(tv))$$
$$= \frac{1}{t}(t \cdot Df_a(v) + \varphi(tv)) = Df_a(v) + \frac{1}{t}\varphi(tv),$$

where φ satisfies $\lim_{h\to 0} \frac{1}{\|h\|}\varphi(h) = 0$. Thus,

$$\partial_v f(a) = g_v'(0) = \lim_{t\to 0} \frac{1}{t}(g_v(t) - g_v(0)) = Df_a(v) + \lim_{t\to 0} \frac{1}{t}\varphi(tv).$$

We further compute that

$$\lim_{t\to 0} \left\| \frac{1}{t}\varphi(tv) \right\| = \lim_{t\to 0} \frac{\|\varphi(tv)\|}{|t|} = \lim_{t\to 0} \left\| \frac{1}{\|tv\|}\varphi(tv) \right\| \cdot \|v\| = 0.$$

By the properties of a norm, this shows $\lim_{t\to 0} \frac{1}{t}\varphi(tv) = 0$, proving the claim. □

Theorem 12.15 implies in particular that *every differentiable map $f : U \to \mathbb{R}^m$ is partially differentiable.* However, the converse statement of Theorem 12.15 is not true. A partially differentiable map does in general *not* need to be differentiable as the following counterexamples show.

Example 12.16. (1) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function from Example 12.6.(2), which we have shown to be partially differentiable. If f was differentiable in $(0,0)$, all directional derivatives would exist by Theorem 12.15, but we have already shown in Example 12.6.(2) that $\partial_{(1,1)} f(0,0)$ does not exist. Hence, f is partially differentiable, but not differentiable.

(2) Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x_1, x_2) = \begin{cases} \frac{x_1^3}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } (x_1, x_2) = (0,0). \end{cases}$$

We want to show that all directional derivatives of f exist in $(0,0)$, but that f is still not differentiable in $(0,0)$. For $v = (v_1, v_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$ we compute for all $t \neq 0$ that

$$\frac{f((0,0) + tv) - f(0,0)}{t} = \frac{f(tv_1, tv_2)}{t} = \frac{\frac{t^3 v_1^3}{t^2 v_1^2 + t^2 v_2^2}}{t} = \frac{v_1^3}{v_1^2 + v_2^2}.$$

Taking limits then shows that

$$\partial_v f(0,0) = \frac{v_1^3}{v_1^2 + v_2^2}, \tag{12.1}$$

so, indeed, all directional derivatives of f exist in $(0,0)$. If f was differentiable in $(0,0)$, then by Theorem 12.15 and the linearity of the differential in $(0,0)$,

$$\partial_{v+w} f(0,0) = Df_{(0,0)}(v+w) = Df_{(0,0)}(v) + Df_{(0,0)}(w) = \partial_v f(0,0) + \partial_w f(0,0) \qquad \forall v, w \in \mathbb{R}^2.$$

(Here we let $\partial_0 f(0,0) := 0$.) But this equation is not valid for our choice of f: we derive from (12.1) that

$$\partial_{(1,1)} f(0,0) = \frac{1}{2}, \quad \partial_{(1,0)} f(0,0) = 1, \quad \partial_{(0,1)} f(0,0) = 0,$$

such that $\partial_{(1,1)} f(0,0) \neq \partial_{(1,0)} f(0,0) + \partial_{(0,1)} f(0,0)$. Hence, f is *not* differentiable in $(0,0)$.

(3) There are examples showing that even if all directional derivatives of f in a point a exist and if the expression $\mathbb{R}^n \to \mathbb{R}^m$, $v \mapsto \partial_v f(a)$, is linear, f might still be not differentiable in a.[2]

The connection between the differential of a map at a point and its directional derivatives can now be used to express the differential in terms of partial derivatives, thereby showing us a possibility of how to explicitly compute differentials.

Corollary 12.17. *Let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be differentiable in $a \in U$. Then*

$$Df_a(v) = \sum_{i=1}^{n} \partial_i f(a) v_i \qquad \forall v = (v_1, \ldots, v_n) \in \mathbb{R}^n.$$

[2]Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x,y) = \begin{cases} x & \text{if } y = x^2, \\ 0 & \text{if } y \neq x^2. \end{cases}$$

One computes that $\partial_v f(0,0) = 0$ for all $v \in \mathbb{R}^2$, but that the zero map (which is the only possible candidate here) does not fulfill the properties of a differential.

Proof. By Theorem 12.15, it particularly holds that $Df_a(e_i) = \partial_i f(a)$ for all $i \in \{1, 2, \dots, n\}$. Since Df_a is linear, we obtain from writing $v = (v_1, \dots, v_n)$ as $v = \sum_{i=1}^n v_i e_i$ that

$$Df_a(v) = Df_a\Big(\sum_{i=1}^n v_i e_i\Big) = \sum_{i=1}^n v_i Df_a(e_i) = \sum_{i=1}^n \partial_i f(a) v_i.$$

□

We have seen in the previous section that partial derivatives can be computed by our well-known one-dimensional rules of differentiation, so with this in mind, Corollary 12.17 shows us how to compute differentials explicitly. But in order to do so, we already need to know that a map is differentiable. While it is not enough to show that a map is partially differentiable, the following theorem, which will be our main criterion to show that a map is differentiable, shows that we can still conclude that a map is differentiable if its partial derivatives are "well-behaved".

Theorem 12.18. *Let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ and $a \in U$. If f is partially differentiable and if the partial derivatives $\partial_i f : U \to \mathbb{R}^m$ are continuous in a for all $i \in \{1, 2, \dots, n\}$, then f is differentiable in a.*

Proof. We first consider the case of $m = 1$, i.e. of $f : U \to \mathbb{R}$. We define $A \in (\mathbb{R}^n)^* = L(\mathbb{R}^n, \mathbb{R})$ by

$$A(h) = \sum_{i=1}^n \partial_i f(a) h_i \quad \forall h = (h_1, \dots, h_n) \in \mathbb{R}^n.$$

We want to show that f is differentiable in a with $Df_a = A$, for which we need to show that

$$\lim_{h \to 0} \frac{|f(a+h) - f(a) - A(h)|}{\|h\|} \stackrel{!}{=} 0.$$

Since U is open, there exists an $\varepsilon > 0$ with $B_\varepsilon(a) \subset U$. Let $h = (h_1, \dots, h_n) \in B_\varepsilon(a)$ and consider $a_0, a_1, \dots, a_n \in \mathbb{R}^n$ given by $a_0 := a$ and

$$a_0 := a, \quad a_j := a_{j-1} + h_j e_j \quad \forall j \in \{1, 2, \dots, n\},$$

such that $a_n = a + h$. One checks that $a_j \in B_\varepsilon(a) \subset U$ for all $j \in \{0, 1, \dots, n\}$ Then:

$$\begin{aligned} f(a+h) - f(a) &= f(a_n) - f(a_{n-1}) + f(a_{n-1}) - f(a_{n-2}) + \cdots + f(a_1) - f(a_0) \\ &= \sum_{j=1}^n (f(a_i) - f(a_{i-1})). \end{aligned}$$

For $j \in \{1, 2, \dots, n\}$ we define the compact interval I_j by

$$I_j := \begin{cases} [0, h_j] & \text{if } h_j \geq 0, \\ [h_j, 0] & \text{if } h_j < 0, \end{cases}$$

and consider $g_j : I_j \to \mathbb{R}$, $g_j(t) = f(a_{j-1} + te_j)$. Then

$$g_j(0) = a_{j-1} \qquad \text{and} \qquad g_j(h_j) = a_j \qquad \forall j \in \{1, 2, \dots, n\}.$$

One checks that each g_j is well-defined, since $a_{j-1} + te_j \in B_\varepsilon(a)$ for all $t \in I_j$, $j \in \{1, 2, \dots, n\}$. Moreover, since f is partially differentiable it follows by construction that each g_j is differentiable with $g_j'(t) = \partial_j f(a_{j-1} + te_j)$ for all $t \in I_j$. By the mean value theorem from Mathematics 1 (Theorem 4.20) there exists a $t_j \in I_j$, such that

$$g_j(h_j) - g_j(0) = g_j'(t_j)(h_j - 0) \quad \Leftrightarrow \quad f(a_j) - f(a_{j-1}) = \partial_j f(c_j) h_j \quad \forall j \in \{1, 2, \dots, n\},$$

where $c_j := a_{j-1} + t_j h_j$. Using these equations, we compute

$$\begin{aligned}|f(a+h) - f(a) - A(h)| &= \left|\sum_{j=1}^{n}(f(a_j) - f(a_{j-1})) - A(h)\right| \\ &= \left|\sum_{j=1}^{n}(\partial_j f(c_j) - \partial_j f(a))h_j\right| \leq \sum_{j=1}^{n}\left|\partial_j f(c_j) - \partial_j f(a)\right| \|h\|,\end{aligned}$$

hence

$$\lim_{h\to 0}\frac{|f(a+h) - f(a) - A(h)|}{\|h\|} \leq \sum_{j=1}^{n}\lim_{h\to 0}\left|\partial_j f(c_j) - \partial_j f(a)\right|,$$

the values of $c_1, \ldots, c_n$ depend on h. But one checks from their definition that if h tends to 0, the c_j will tend to 0 as well. So it follows from the continuity of the partial derivatives in a and the squeeze theorem for functions that

$$\lim_{h\to 0}\frac{|f(a+h) - f(a) - A(h)|}{\|h\|} = 0,$$

which proves the claim.
For $m > 1$, i.e. for $f = (f_1, \ldots, f_m)$, where $f_i : U \to \mathbb{R}$ for all $i \in \{1, 2, \ldots, m\}$, we observe that by definition

$$\partial_j f(a) = (\partial_j f_1(a), \ldots, \partial_j f_m(a)) \quad \forall j \in \{1, 2, \ldots, n\}.$$

Hence if $\partial_j f$ is continuous in a for all j, then by Theorem 11.41 $\partial_j f_i$ is continuous in a for all $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$, so it follows from the case of $m = 1$ that f_i is differentiable in a for all $i \in \{1, 2, \ldots, m\}$. By Theorem 12.14, this shows that f is differentiable in a. □

Example 12.19. (1) We have seen in Example 12.9.(2) that

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \qquad f(x_1, x_2) = (e^{-x_1^2 - x_2^2}, \sin(x_1 x_2)),$$

is partially differentiable with

$$\partial_1 f(x_1, x_2) = (-2x_1 e^{-x_1^2 - x_2^2}, x_2\cos(x_1 x_2)), \quad \partial_2 f(x_1, x_2) = (-2x_2 e^{-x_1^2 - x_2^2}, x_1\cos(x_1 x_2)).$$

for all $(x_1, x_2) \in \mathbb{R}^2$. These partial derivatives are componentwise compositions of continuous maps, hence themselves continuous, so it follows from Theorem 12.18 that f is differentiable. By Corollary 12.17 its differential in $(x_1, x_2) \in \mathbb{R}^2$ is given by

$$\begin{aligned}Df_{(x_1,x_2)}(v_1, v_2) &= \partial_1 f(x_1, x_2)v_1 + \partial_2 f(x_1, x_2)v_2 \\ &= (-2x_1 e^{-x_1^2 - x_2^2}, x_2\cos(x_1 x_2)) \cdot v_1 + (-2x_2 e^{-x_1^2 - x_2^2}, x_1\cos(x_1 x_2)) \cdot v_2 \\ &= (-2e^{-x_1^2 - x_2^2}(x_1 v_1 + x_2 v_2), \cos(x_1 x_2)(x_2 v_1 + x_1 v_2))\end{aligned}$$

for all $(v_1, v_2) \in \mathbb{R}^2$.

(2) The converse of Theorem 12.18 is fals in general, i.e. there are differentiable maps whose partial derivatives are not continuous. As a counterexample, it suffices to find a function $f : I \to \mathbb{R}$, where I is an interval, which is differentiable and whose first (and only) partial derivative, which is nothing but $\partial_1 f = f'$ is not continuous. As the reader may verify, such an example is given by

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} x^2\sin(\frac{1}{x}) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Corollary 12.20. *Every polynomial $p : \mathbb{R}^n \to \mathbb{R}$, where $n \in \mathbb{N}$, is differentiable.*

Proof. Apparently, each polynomial is partially differentiable and since the partial derivatives of a polynomial are again polynomials, which are continuous, this follows from Theorem 12.18. □

Remark 12.21. In this section, we have established the following implications about maps of the form $f : U \to \mathbb{R}^m$, where $U \subset \mathbb{R}^n$ is open:

f is partially differentiable and all partial derivatives are continuous in $a \in U$

$\Downarrow$

f is differentiable in a

$\Downarrow$

all directional derivatives of f exists in a

$\Downarrow$

f is partially differentiable in a

Moreover, *none of the converse implications holds*. A counterexample for the converse of the third implication is given in Example 12.6.(2), for the converse of the second implication in Example 12.16.(2) and for the first implication in Example 12.19.(2).

12.3 Rules of multivariable differentiation

Similar to the one-dimensional case, there are several rules and methods that allow us to compute differentials explicitly and we shall discuss them in this section. Not only are there generalizations of the product rule and the chain rule to higher dimensions, but also can we use our knowledge about linear algebra and matrices to compute compositions of differentials by matrix multiplication.
Before we carry out the details, we first study the differentiability of one of our favorite classes of maps, namely linear maps. Since we interpreted the differential of a map in a point as the best approximation of the map by a linear one, we already have a guess of what the best linear approximation of a linear map is: the map itself. This is confirmed in the next theorem.

Theorem 12.22. *Let $m, n \in \mathbb{N}$ and let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a* linear *map. Then f is differentiable with*

$$DL_a = L \qquad \forall a \in \mathbb{R}^n.$$

Proof. Let $a \in \mathbb{R}^n$. By linearity, it holds for all $h \in \mathbb{R}^n$ that

$$L(a + h) = L(a) + L(h).$$

Taking a look at the defining equation of the differential, we see that L satisfies the equation with respect to $\varphi(h) = 0$ for all $h \in \mathbb{R}^n$. Thus, $DL_a = L$. □

The following mildly surprising result shows that differentials behave well with respect to sums and multiplication by scalars.

Theorem 12.23. *Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $a \in U$. Let $f, g : U \to \mathbb{R}^m$ both be differentiable in a. Then:*

a) $f + g : U \to \mathbb{R}^m$ is differentiable in a with $D(f + g)_a = Df_a + Dg_a$.

b) $\lambda f : U \to \mathbb{R}^m$ *is differentiable in* a *with* $D(\lambda f)_a = \lambda Df_a$ *for all* $\lambda \in \mathbb{R}$.

Proof. We will only prove part a), since part b) is proven by a very similar argument. Since both f and g are differentiable in a, there are $\varphi_1, \varphi_2 : U' \to \mathbb{R}^m$, where $U' = \{h \in \mathbb{R}^n \mid a + h \in U\}$, with $\lim_{h\to 0} \frac{1}{\|h\|}\varphi_i(h) = 0$ for $i \in \{1,2\}$, such that

$$f(a+h) = f(a) + Df_a(h) + \varphi_1(x), \qquad g(a+h) = g(a) + Dg_a(h) + \varphi_2(x) \quad \forall h \in U'.$$

Adding these two equations yields

$$\begin{aligned}(f+g)(a+h) &= f(a+h) + g(a+h)\\ &= f(a) + g(a) + Df_a(h) + Dg_a(h) + \varphi_1(h) + \varphi_2(h)\\ &= (f+g)(a) + (Df_a + Dg_a)(h) + \varphi(h),\end{aligned}$$

where $\varphi := \varphi_1 + \varphi_2$ satisfies $\lim_{h\to 0} \frac{1}{\|h\|}\varphi(h) = 0$. Since $Df_a + Dg_a$ is again a linear map, this shows that $f + g$ is differentiable in a with $D(f+g)_a = Df_a + Dg_a$. □

We next want to establish a multivariable version of the product rule. Note that in contrast to the previous theorem, products of maps are only well-defined for maps whose codomain is $\mathbb{R}$, for which we can multiply the values of the functions as real numbers.

Theorem 12.24 (product rule)**.** *Let* $n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ *open and* $a \in U$. *If* $f, g : U \to \mathbb{R}$ *are both differentiable in* a, *then* $f \cdot g : U \to \mathbb{R}$ *is differentiable in* a *with*

$$D(f \cdot g)_a = Df_a \cdot g(a) + f(a) \cdot Dg_a,$$

i.e. $D(f \cdot g)_a(v) = Df_a(v) \cdot g(a) + f(a) \cdot Dg_a(v)$ *for all* $v \in \mathbb{R}^n$.

Proof. We want to show the claim by showing that

$$\lim_{h\to 0} \frac{|(f\cdot g)(a+h) - (f\cdot g)(a) - Df_a(h)\cdot g(a) - f(a)Dg_a(h)|}{\|h\|} \overset{!}{=} 0.$$

By assumption, there are $\varphi_1, \varphi_2 : U' \to \mathbb{R}$ with $\lim_{h\to 0} \frac{\varphi_i(h)}{\|h\|} = 0$ for $i \in \{1,2\}$, where $U' = \{h \in \mathbb{R}^n \mid a + h \in U\}$, such that

$$f(a+h) = f(a) + Df_a(h) + \varphi_1(x), \qquad g(a+h) = g(a) + Dg_a(h) + \varphi_2(x) \quad \forall h \in U'.$$

Taking the product of these two equations we obtain

$$\begin{aligned}(f\cdot g)(a+h) &= f(a+h)\cdot g(a+h)\\ &= f(a)\cdot g(a) + f(a)\cdot Dg_a(h) + Df_a(h)\cdot g(a) + Df_a(h)\cdot Dg_a(h)\\ &\quad + \varphi_1(h)(g(a) + Dg_a(h)) + (f(a) + Df_a(h))\varphi_2(h) + \varphi_1(h)\cdot\varphi_2(h)\end{aligned}$$

(The rest of the proof is not discussed in the lecture videos.) By Theorem 11.44 and Corollary 11.42, there are $C_1, C_2 \in (0, +\infty)$, such that $|Df_a(v)| \le C_1\|v\|$ and $|Dg_a(v)| \le C_2\|v\|$ for all $v \in \mathbb{R}^n$. Using this fact, we derive from the previous computation for all $h \in U'$:

$$\begin{aligned}&|(f\cdot g)(a+h) - (f\cdot g)(a) - Df_a(h)\cdot g(a) - f(a)Dg_a(h)|\\ &= |Df_a(h)\cdot Dg_a(h) + \varphi_1(h)(g(a) + Dg_a(h)) + (f(a) + Df_a(h))\varphi_2(h) + \varphi_1(h)\cdot\varphi_2(h)|\\ &\le |Df_a(h)|\cdot|Dg_a(h)||\varphi_1(h)| \cdot |g(a) + Dg_a(h)| + |f(a) + Df_a(h)|\cdot|\varphi_2(h)| + |\varphi_1(h)|\cdot|\varphi_2(h)|\\ &\le C_1C_2\|h\|^2 + |\varphi_1(h)|\cdot|g(a) + Dg_a(h)| + |f(a) + Df_a(h)|\cdot|\varphi_2(h)| + |\varphi_1(h)|\cdot|\varphi_2(h)|.\end{aligned}$$

This yields:

$$\begin{aligned}
&\lim_{h\to 0} \frac{|(f\cdot g)(a+h) - (f\cdot g)(a) - Df_a(h)\cdot g(a) - f(a)Dg_a(h)|}{\|h\|} \\
&\leq \lim_{h\to 0} \left(C_1C_2\|h\| + \frac{|\varphi_1(h)|}{\|h\|}|g(a) + Dg_a(h)| + |f(a) + Df_a(h)|\frac{|\varphi_2(h)|}{\|h\|} + \frac{|\varphi_1(h)|}{\|h\|}\cdot\frac{|\varphi_2(h)|}{\|h\|}\cdot\|h\| \right) \\
&= 0 + 0\cdot|g(a)| + |f(a)|\cdot 0 + 0\cdot 0\cdot 0 = 0.
\end{aligned}$$

This shows the claim. □

Remark 12.25. In the situation of Theorem 12.24, inserting the canonical basis vectors into the differentials yields the following formula for the corresponding partial derivatives:

$$(\partial_i(f\cdot g))(a) = \partial_i f(a)g(a) + f(a)\partial_i g(a) \qquad \forall i \in \{1,2,\dots,n\}.$$

Example 12.26. Let $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = e^{-x^2-y^2}\sin(xy)$. Then $f = g\cdot h$, where $g(x,y) = e^{-x^2-y^2}$ and $h(x,y) = \sin(x,y)$. Both g and h are differentiable, so by Theorem 12.24, f is differentiable. From the formula from Remark 12.25 we obtain for all $(x,y) \in \mathbb{R}^2$:

$$\begin{aligned}
\partial_1 f(x,y) &= \partial_1 g(x,y)h(x,y) + g(x,y)\partial_1 h(x,y) \\
&= -2xe^{-x^2-y^2}\sin(xy) + e^{-x^2-y^2}y\cos(xy) = e^{-x^2-y^2}(-2x\sin(xy) + y\cos(xy)), \\
\partial_2 f(x,y) &= \partial_2 g(x,y)h(x,y) + g(x,y)\partial_2 h(x,y) \\
&= -2ye^{-x^2-y^2}\sin(xy) + e^{-x^2-y^2}x\cos(xy) = e^{-x^2-y^2}(-2y\sin(xy) + yx\cos(xy)).
\end{aligned}$$

Hence, for all $(x,y),(v_1,v_2) \in \mathbb{R}^2$ we derive that

$$Df_{(x,y)}(v_1,v_2) = e^{-x^2-y^2}((-2x\sin(xy) + y\cos(xy))v_1 + (-2y\sin(xy) + yx\cos(xy))v_2).$$

Next we establish a chain rule for compositions of differentiable maps in several variables. Together with the rules that we have established so far, this enables us to show the differentiability of many different maps and to compute their derivatives.

Theorem 12.27 (chain rule)**.** *Let $\ell, m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open, let $f : U \to V$ and $g : V \to \mathbb{R}^\ell$ and let $a \in U$. If f is differentiable in a and g is differentiable in $f(a)$, then $g\circ f : U \to \mathbb{R}^\ell$ is differentiable in a with*

$$D(g\circ f)_a = Dg_{f(a)} \circ Df_a.$$

Proof. To show the claim, it suffices to show that

$$\lim_{h\to 0} \frac{\|(g\circ f)(a+h) - (g\circ f)(a) - (Dg_{f(a)}\circ Df_a)(h)\|}{\|h\|} \overset{!}{=} 0.$$

Since g is differentiable in $f(a)$, there exists $\varphi_2 : V' \to \mathbb{R}^\ell$, where $V' = \{k \in \mathbb{R}^m \mid f(a)+k \in V\}$, with $\lim_{k\to 0}\frac{1}{\|k\|}\varphi_2(k) = 0$, such that

$$g(a+k) = g(a) + Dg_a(k) + \varphi_2(k) \qquad \forall k \in V'.$$

In particular, we compute for suitable $h \in \mathbb{R}^n$ that

$$\begin{aligned}
(g\circ f)(a+h) &= g(f(a+h)) \\
&= g(f(a) + (f(a+h) - f(a))) \\
&= g(f(a)) + Dg_{f(a)}(f(a+h) - f(a)) + \varphi_2(f(a+h) - f(a)).
\end{aligned}$$

Since f is differentiable in a, there exists $\varphi_1 : U' \to \mathbb{R}^m$, where $U' = \{h \in \mathbb{R}^n \mid a + h \in U\}$, with $\lim_{h\to 0} \frac{1}{\|h\|}\varphi_1(h) = 0$ so that

$$f(a+h) - f(a) = Df_a(h) + \varphi_1(h) \qquad \forall h \in U'.$$

Inserting this into the above computation yields

$$\begin{aligned}(g \circ f)(a+h) &= g(f(a)) + Dg_{f(a)}(Df_a(h) + \varphi_1(h)) + \varphi_2(Df_a(h) + \varphi_1(h)) \\ &= (g \circ f)(a) + Dg_{f(a)}(Df_a(h)) + Dg_{f(a)}(\varphi_1(h)) + \varphi_2(Df_a(h) + \varphi_1(h)),\end{aligned}$$

where we used the linearity of $Dg_{f(a)}$. By Corollary 11.42 and Theorem 11.44 there are $C_1, C_2 > 0$, such that $\|Df_a(v)\| \le C_1\|v\|$ for all $v \in \mathbb{R}^n$ and $\|Dg_{f(a)}(w)\| \le C_2\|w\|$ for all $w \in \mathbb{R}^m$. Using this and the above computation, we obtain for all $h \in U' \setminus \{0\}$:

$$\begin{aligned}\|(g \circ f)(a+h) - (g \circ f)(a) - (Dg_{f(a)} \circ Df_a)(h)\| &= \|Dg_{f(a)}(\varphi_1(h) + \varphi_2(Df_a(h) + \varphi_1(h))\| \\ &\le \|Dg_{f(a)}(\varphi_1(h))\| + \|\varphi_2(Df_a(h) + \varphi_1(h))\| \\ &\le C_2\|\varphi_1(h)\| + \|\varphi_2(Df_a(h) + \varphi_1(h))\|.\end{aligned}$$

Consequently,

$$\begin{aligned}&\lim_{h\to 0} \frac{\|(g \circ f)(a+h) - (g \circ f)(a) - (Dg_{f(a)} \circ Df_a)(h)\|}{\|h\|} \\ &\le C_2 \lim_{h\to 0} \frac{\|\varphi_1(h)\|}{\|h\|} + \lim_{h\to 0} \frac{\|\varphi_2(Df_a(h) + \varphi_1(h))\|}{\|h\|} = \lim_{h\to 0} \frac{\|\varphi_2(Df_a(h) + \varphi_1(h))\|}{\|h\|}.\end{aligned}$$

It only remains to show that this latter limit vanishes. Since $\lim_{h\to 0} \frac{\|\varphi_1(h)\|}{\|h\|} = 0$, it holds in a neighborhood of 0 that $\|\varphi_1(h)\| \le \|h\|$. Hence, for every h in such a neighborhood it holds that

$$\|Df_a(h) + \varphi_1(h)\| \le \|Df_a(h)\| + \|\varphi_1(h)\| \le C_1\|h\| + \|h\| = (C_1 + 1)\|h\|$$

and thus

$$\lim_{h\to 0} \frac{\|\varphi_2(Df_a(h) + \varphi_1(h))\|}{\|h\|} \le (C_1 + 1) \lim_{h\to 0} \frac{\|\varphi_2(Df_a(h) + \varphi_1(h))\|}{\|Df_a(h) + \varphi_1(h))\|} = (C_1 + 1) \lim_{h\to 0} \frac{\|\varphi_2(h)\|}{\|h\|} = 0.$$

This shows the claim. □

Example 12.28. (1) Let $h : \mathbb{R} \to \mathbb{R}$ be continuous and let

$$F : \mathbb{R}^2 \to \mathbb{R}, \qquad F(x,y) = \int_0^{\sin(x+y)} h(t)\, dt.$$

Then $F = g \circ f$, where $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = \sin(x+y)$, and $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \int_0^x h(t)\, dt$. The map f has continuous partial derivatives, hence is differentiable by Theorem 12.18 with

$$Df_{(x,y}(v_1, v_2) = \partial_1 f(x,y)v_1 + \partial_2 f(x,y)v_2 = \cos(x+y)(v_1 + v_2) \qquad \forall (x,y), (v_1, v_2) \in \mathbb{R}^2.$$

By the fundamental theorem of calculus, g is differentiable with $g'(x) = h(x)$ for all $x \in \mathbb{R}$. Hence, F is differentiable by Theorem 12.27 and its differential at $(x,y) \in \mathbb{R}^2$ is given by

$$\begin{aligned}DF_{(x,y)}(v_1, v_2) &= Dg_{f(x,y)}(Df_{(x,y)}(v_1, v_2)) = g'(f(x,y)) \cdot Df_{(x,y)}(v_1, v_2) \\ &= h(\sin(x+y))\cos(x+y)(v_1 + v_2). \qquad \forall (v_1, v_2) \in \mathbb{R}^2.\end{aligned}$$

(2) *(This example is not carried out in detail in the lecture videos.)*

Consider $h := g \circ f : \mathbb{R}^2 \to \mathbb{R}$, where

$$f : \mathbb{R}^2 \to \mathbb{R}^3,\ f(x,y) = (e^x \sin y, e^x \cos y, e^x), \qquad g : \mathbb{R}^3 \to \mathbb{R},\ g(x,y,z) = 3x^2 - 2xy + 4z^2.$$

Both f and g are partially differentiable. We compute for all $(x,y) \in \mathbb{R}^2$ that

$$\begin{aligned} &\partial_1 f(x,y) = (e^x \sin y, e^x \cos y, e^x), \qquad \partial_2 f(x,y) = (e^x \cos y, -e^x \sin y, 0), \\ &\partial_1 g(x,y,z) = 6x - 2y, \qquad \partial_2 g(x,y,z) = -2x, \qquad \partial_3 g(x,y,z) = 8z. \end{aligned}$$

All of these derivatives are obviously continuous. Hence, f and g are differentiable by Theorem 12.18 and so is h by Theorem 12.27. We compute using Corollary 12.17 that

$$\begin{aligned} Df_{(x,y)}(v_1,v_2) &= \partial_1 f(x,y)v_1 + \partial_2 f(x,y)v_2, \\ &= (e^x \sin(y)v_1, e^x \cos(y)v_1, e^x v_1) + (e^x \cos(y)v_2, -e^x \sin(y)v_2, 0) \\ &= (e^x(\sin(y)v_1 + \cos(y)v_2), e^x(\cos(y)v_1 - \sin(y)v_2), e^x v_1). \end{aligned}$$

for all $(v_1,v_2) \in \mathbb{R}^2$. Thus, by Theorem 12.27 it holds for all $(x,y),(v_1,v_2) \in \mathbb{R}^2$ that

$$\begin{aligned} &Dh_{(x,y)}(v_1,v_2) = D(g \circ f)_{(x,y)}(v_1,v_2) = Dg_{f(x,y)}(Df_{(x,y)}(v_1,v_2)) \\ &= Dg_{(e^x \sin y, e^x \cos y, e^x)}(e^x(\sin(y)v_1 + \cos(y)v_2), e^x(\cos(y)v_1 - \sin(y)v_2), e^x v_1) \\ &= \partial_1 g(e^x \sin y, e^x \cos y, e^x)e^x(\sin(y)v_1 + \cos(y)v_2) \\ &\quad + \partial_2 g(e^x \sin y, e^x \cos y, e^x)e^x(\cos(y)v_1 - \sin(y)v_2) + \partial_3 g(e^x \sin y, e^x \cos y, e^x)e^x v_1 \\ &= (6e^x \sin y - 2e^x \cos y)e^x(\sin(y)v_1 + \cos(y)v_2) - 2e^x \sin(y)e^x(\cos(y)v_1 - \sin(y)v_2) + 8(e^x)^2 v_1 \\ &= e^{2x}(6\sin^2(y) - 4\sin y \cos y + 8)v_1 + e^{2x}(6\sin y \cos y - 2\cos^2(y) + 2\sin^2(y))v_2 \\ &(= e^{2x}(3\cos(2y) - 2\sin(2y) + 11)v_1 + e^{2x} + e^{2x}(3\sin(2y) - 2\cos(2y))v_2), \end{aligned}$$

where we used several trigonometric identities in the last equation.

There is one special case of the chain rule that is worth pointing out and formulating as a separate statement.

Corollary 12.29. *Let I be an interval, $U \subset \mathbb{R}^n$ be open, $\gamma : I \to U$ be a curve in $\mathbb{R}^n$, $f : U \to \mathbb{R}$ and $t_0 \in I$. If $\gamma = (\gamma_1, \ldots, \gamma_n)$ is differentiable in t_0 and f is differentiable in $\gamma(t_0)$, then $f \circ \gamma : I \to \mathbb{R}$ is differentiable in t_0 with*

$$(f \circ \gamma)'(t_0) = Df_{\gamma(t_0)}(\gamma'(t_0)) = \sum_{i=1}^{n} \partial_i f(\gamma(t_0))\gamma_i'(t_0).$$

Proof. By Theorem 4.6, the relation between the derivative of the one-dimensional real function $f \circ \gamma$ and its differential is given by

$$D(f \circ \gamma)_{t_0}(h) = (f \circ \gamma)'(t_0) \cdot g \quad \forall h \in \mathbb{R}.$$

In particular, by Theorem 12.27,

$$\begin{aligned} (f \circ \gamma)'(t_0) &= D(f \circ \gamma)_{t_0}(1) = Df_{\gamma(t_0)}(D\gamma_{t_0}(1)) \\ &= Df_{\gamma(t_0)}(\gamma'(t_0)) = \sum_{i=1}^{n} \partial_i f(\gamma(t_0))\gamma_i'(t_0), \end{aligned}$$

where we have used Corollary 12.17. □

As differentials of maps are linear maps of the form $\mathbb{R}^n \to \mathbb{R}^m$, we know from Section 8.1 that each such map is induced by a real $m \times n$-matrix. We determine the entries of this matrix explicitly in the next statement.

Theorem/Definition 12.30. *Let $m, n \in \mathbb{R}$, let $U \subset \mathbb{R}^n$ be open, let $a \in U$ and let $f : U \to \mathbb{R}^m$ be differentiable in a. Then the linear map $Df_a : \mathbb{R}^n \to \mathbb{R}^m$ is induced by the matrix*

$$J_f(a) \in M(m \times n, \mathbb{R}), \qquad J_f(a) = (\partial_j f_i(a)) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \dots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \dots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \dots & \partial_n f_m(a) \end{pmatrix}.$$

$J_f(a)$ is called the Jacobi matrix of f in a.

Proof. As we already observed in Remark 8.6 for arbitrary linear maps $\mathbb{R}^n \to \mathbb{R}^m$, the linear map Df_a is induced by the matrix $A \in M(m \times n, \mathbb{R})$ whose j-th column is the image of the canonical basis vector e_j under Df_a. But by Corollary 12.17,

$$Df_a(e_j) = \partial_j f(a) \cdot 1 = (\partial_j f_1(a), \partial_j f_2(a), \dots, \partial_j f_m(a)).$$

Identifying this with a column vector and assembling these columns to a matrix, one sees that indeed $A = J_f(a)$ as defined in the statement of the theorem. □

Corollary 12.31. *Let $\ell, m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open, let $f : U \to V$ and $g : V \to \mathbb{R}^\ell$ and let $a \in U$. If f is differentiable in a and g is differentiable in $f(a)$, then*

$$J_{g \circ f}(a) = J_g(f(a)) \cdot J_f(a),$$

where $\cdot$ denotes the matrix product. Componentwise, if $f = (f_1, \dots, f_m)$, $g = (g_1, \dots, g_\ell)$ and $g \circ f = (h_1, \dots, h_\ell)$, then

$$\partial_j h_i(a) = \sum_{k=1}^{m} \partial_j f_k(a) \partial_k g_i(a) \qquad \forall i \in \{1, 2, \dots, \ell\},\ j \in \{1, 2, \dots, n\}.$$

Proof. This is an immediate consequence of Theorem 12.27 and Theorem/Definition 8.9 about matrices and compositions of linear maps. □

12.4 Further properties of differentiable maps

In this section we collect several important results about differentiable maps and make our first steps towards higher order differentiability for maps in several variables. We begin by discussing properties of differentiable real functions, i.e. differentiable maps $f : U \to \mathbb{R}$, where $U \subset \mathbb{R}^n$ for some $n \in \mathbb{N}$. In this case, the Jacobi matrix of f in some $a \in U$ is by definition a real $1 \times n$-matrix, i.e. a row vector. We identify this row vector with an element of $\mathbb{R}^n$ and explore its properties in the first part of this section.

Definition 12.32. Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$. If f is partially differentiable in $a \in U$, then we call[3]

$$\nabla f(a) := (\partial_1 f(a), \partial_2 f(a), \dots, \partial_n f(a)) \in \mathbb{R}^n$$

[3]The symbol ∇ is called "nabla" and is not a letter of the Greek or any other alphabet. It was introduced without a name by William R. Hamilton and later named nabla, which is the Greek name for a Phoenician harp mentioned in the bible which is supposed to have had a similar shape. (Yes, this is a true story.)

the *gradient of f in a*. If f is partially differentiable, then the map $\nabla f : U \to \mathbb{R}^n$ is called the *gradient of f* or *gradient (vector) field of f*. (Other common notation: grad f.)

The immediate connection between gradient and differential is given by the following statement.

Corollary 12.33. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable in $a \in U$. Then*

$$Df_a(v) = \partial_v f(a) = \langle \nabla f(a), v \rangle \qquad \forall v \in \mathbb{R}^n \setminus \{0\},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$.

Proof. This follows immediately from Corollary 12.17 and the definition of the standard inner product. □

Remark 12.34. (1) In the situation of the Corollary 12.33, the mere existence of a vector $v_0 \in \mathbb{R}^n$ which satisfies $Df_a(v) = \langle v_0, v \rangle$ for all $v \in V$ follows already from the Riesz representation theorem (Theorem 10.28.c)). We might also have defined the gradient as that vector predicted by Theorem 10.28.c) and then computed from this definition, that the entries of the vector must be the partial derivatives of f.

(2) Let $v \in \mathbb{R}^n$ with $\|v\| = 1$. Then

$$\partial_v f(a) = \langle \nabla f(a), v \rangle = \|\nabla f(a)\| \cdot \cos(\angle(\nabla f(a), v)).$$

Assume that $\nabla f(a) \neq 0$. Then the above equation shows that $\partial_v(a)$ becomes maximal if and only if $\angle(\nabla f(a), v) = 0$, i.e. if $\nabla f(a)$ and v point in the same direction. This shows a crucial geometric property of the gradient: *it points to the direction, in which the function has the "fastest increase"*. Since then $\|\nabla f(a)\| = \partial_v f(a)$, *the length of the gradient is the rate of increase in that direction*. These properties are of great use in various fields of mathematics, physics and engineering, e.g. for gradient-descent methods in optimization problems.

We will use the gradient to find an generalization of Fermat's theorem from Mathematics 1 (Theorem 4.17) to functions in several real variables, i.e. a necessary condition on such a function to have a local maximum or a local minimum in a point To do so, we first give a formal definition of local maxima and minima in our higher-dimensional setting. The reader might notice that the definition is completely analogous to local extreme of functions defined on subsets of $\mathbb{R}$ or $\mathbb{C}$ which we studied in Mathematics 1.

Definition 12.35. Let $n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ and $x_0 \in U$. We say that *f has a local maximum (local minimum) in x_0* if there exists a neighborhood A of x_0 with $A \subset U$, such that

$$f(x) \leq f(x_0) \quad \forall x \in A \qquad (f(x) \geq f(x_0) \quad \forall x \in A).$$

the value $f(x_0) \in \mathbb{R}$ is then called a *local maximum (local minimum)* and x_0 is called a *local maximum point (local minimum point) of f*. We further say that *f has a local extremum in x_0* if it has a local maximum in x_0 or a local minimum in x_0.

Theorem 12.36. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be partially differentiable. If f has a local maximum or a local minimum in $a \in U$, then*

$$\nabla f(a) = 0.$$

Proof. We assume w.l.o.g. that f has a local maximum in a and let $\varepsilon > 0$, such that $B_\varepsilon(a) \subset U$ and $f(x) \leq f(a)$ for all $x \in B_\varepsilon(a)$. We need to show that $\partial_i f(a) = 0$ for each $i \in \{1, 2, \ldots, n\}$. For each $i \in \{1, 2, \ldots, n\}$ we consider the well-defined function

$$g_i : (-\varepsilon, \varepsilon) \to \mathbb{R}, \qquad g_i(t) = f(a + te_i).$$

Since f is partially differentiable, g_i is differentiable with $g_i'(0) = \partial_i f(a)$ for all $i \in \{1, 2, \ldots, n\}$. By the local maximum property of f, it follows in particular for each i that $g_i(t) = f(a + te_i) \leq f(a) = g_i(0)$ for all $t \in (-\varepsilon, \varepsilon)$, such that g_i has a local maximum in 0. But then by Fermat's theorem (Theorem 4.17) from Mathematics 1, it holds that $g_i'(0) = \partial_i f(a) = 0$ for all $i \in \{1, 2, \ldots, n\}$. □

The mean value theorem for differentiable functions on intervals (Theorem 4.20) has several generalizations to functions in several variables. In this lecture, we will not discuss the most general version, but instead settle with a mean value theorem for maps with values in $\mathbb{R}$, which is proven by reducing it to the one-dimensional result and using the chain rule in a smart way.

Theorem 12.37 (mean value theorem for real functions)**.** *Let $n \in \mathbb{N}$ and let $U \subset \mathbb{R}^n$ be open. Let $f : U \to \mathbb{R}$ be differentiable and let $a, b \in U$ be such that the line segment connecting a and b, i.e. the set $L := \{a + t(b - a) \in \mathbb{R}^n \mid t \in U\}$, lies in U. Then there exists $\xi \in L$, such that*

$$f(b) - f(a) = Df_\xi(b - a).$$

Proof. Let $c : [0, 1] \to U$, $c(t) = a + t(b - a)$. Then c is differentiable with $c'(t) = b - a$ for all $t \in [0, 1]$. By the one-dimensional mean value theorem from Mathematics 1 (Theorem 4.20) and by Corollary 12.29, it follows that there exists $t_0 \in [0, 1]$ with

$$\begin{aligned} f(b) - f(a) &= (f \circ c)(1) - (f \circ c)(0) = (f \circ c)'(t_0) \cdot (1 - 0) \\ &= (f \circ c)'(t_0) = Df_{c(t_0)}(c'(t)) = Df_{c(t_0)}(b - a). \end{aligned}$$

Thus, the claim follows with $\xi := c(t_0)$. □

To apply Theorem 12.37 to large subsets of $\mathbb{R}^n$, we would of course like f to be defined on subsets of $\mathbb{R}^n$, for which the line segment connecting any two points in the set is itself completely contained in the set. Sets with this property have a special name, because they are important in various fields of mathematics, from optimization to geometry.

Definition 12.38. Let $n \in \mathbb{N}$. A subset $K \subset \mathbb{R}^n$ is called *convex* if

$$\{x + t(y - x) \in \mathbb{R}^n \mid t \in [0, 1]\} \subset K \qquad \forall x, y \in K,$$

i.e. if for any two $x, y \in K$ the line segment connecting x and y is entirely contained in K.

Example 12.39. For any $n \in \mathbb{N}$, $a \in \mathbb{R}^n$ and $r > 0$, the open ball $B_r(a)$ and the closed ball $\overline{B}_r(a)$ are convex subsets of $\mathbb{R}^n$. This is shown by a simple computation that is left to the reader.

Corollary 12.40. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and convex and let $f : U \to \mathbb{R}$ be differentiable. If $Df_a = 0$ for all $a \in U$, then f is constant.*

Proof. This follows immediately from Theorem 12.37, since it yields that $f(b) - f(a) = 0$ for all $a, b \in U$. □

The next definition gives a name to a property that we have already studied in Theorem 12.18.

Definition 12.41. Let $m, n \in \mathbb{N}$ and let $U \subset \mathbb{R}^n$ be open. We call $f : U \to \mathbb{R}^m$ *continuously differentiable* if f is partially differentiable and if $\partial_i f : U \to \mathbb{R}^m$ is continuous for every $i \in \{1, 2, \ldots, n\}$.

Recall that by Theorem 12.18, every continuously differentiable function is differentiable.

Theorem 12.42. *Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be continuously differentiable. If $K \subset U$ is compact and convex, then $f|_K$ is Lipschitz-continuous, i.e. there exists $C > 0$, so that*

$$\|f(x) - f(y)\| \leq C \cdot \|x - y\| \qquad \forall x, y \in K.$$

Proof. Let $f = (f_1, \ldots, f_m)$, where $f_i : U \to \mathbb{R}$ for all $i \in \{1, 2, \ldots, m\}$. Since f is continuously differentiable, $\partial_j f_i : U \to \mathbb{R}$ is continuous for all $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, m\}$. By Corollary 11.77, each $\partial_j f_i$ attains its maximum and minimum on K, hence,

$$M_{i,j} := \sup_{x \in K} |\partial_j f_i(x)| < +\infty.$$

Since K is convex, by Theorem 12.37, for all $i \in \{1, 2, \ldots, m\}$ and all $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n) \in K$ there exists $\xi \in K$ with

$$\begin{aligned} |f_i(x) - f_i(y)| = |(Df_i)_\xi(x - y)| &= \Big| \sum_{j=1}^n \partial_j f_i(\xi)(x_j - y_j) \Big| \\ &\leq \sum_{j=1}^n |\partial_j f_i(\xi)||x_j - y_j| \leq \sum_{j=1}^n M_{i,j}|x_j - y_j| \leq M_i \sum_{j=1}^n |x_j - y_j|, \end{aligned}$$

where $M_i := \max\{M_{i,1}, M_{i,2}, \ldots, M_{i,n}\}$ for each i. From this inequality we derive for all $x, y \in K$:

$$\|f(x) - f(y)\|^2 = \sum_{i=1}^m |f_i(x) - f_i(y)|^2 \leq \sum_{i=1}^m M_i^2 \Big(\sum_{j=1}^n |x_j - y_j| \Big)^2. \tag{12.2}$$

We observe that in terms of the standard inner product, we can express and estimate the sum at the end of the previous inequality as follows, using the Cauchy-Schwarz inequality:

$$\begin{aligned} \sum_{j=1}^n |x_j - y_j| &= \langle (|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|), (1, 1, \ldots, 1) \rangle \\ &\leq \|(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|)\| \cdot \|(1, 1, \ldots, 1)\| \\ &= \sqrt{\sum_{j=1}^n (x_j - y_j)^2} \sqrt{n} = \sqrt{n} \|x - y\|. \end{aligned}$$

Inserting this into (12.2) yields

$$\|f(x) - f(y)\|^2 \leq n \sum_{i=1}^m M_i^2 \|x - y\|^2.$$

Thus, with $C := \sqrt{n \sum_{i=1}^m M_i^2}$ we derive that $\|f(x) - f(y)\| \leq C\|x - y\|$ for all $x, y \in K$. □

As for real functions on intervals, we want to consider higher-order differentials and derivatives for maps in several real variables as well. As we did in Section 12.1, we start by discussing partial derivatives of higher order and their properties before discussing higher order differentials in the next section.

Definition 12.43. Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be partially differentiable.

a) We say that f is *twice continuously differentiable* if f is continuously differentiable and if $\partial_j f : U \to \mathbb{R}^m$ is continuously differentiable for every $j \in \{1, 2, \dots, n\}$. The partial derivatives

$$\partial_i \partial_j f := \partial_i(\partial_j f) : U \to \mathbb{R}^m,$$

where $i, j \in \{1, 2, \dots, n\}$, are then called the *second order partial derivatives of* f.

b) Let $k \in \mathbb{N}$ with $k \geq 3$. We say that f is k *times continuously differentiable* if f is $k-1$ timers continuously differentiable and all partial derivatives of f of order $k-1$, i.e. all $\partial_{i_1} \partial_{i_2} \dots \partial_{i_{k-1}} f : U \to \mathbb{R}$, where $i_1, i_2, \dots, i_{k-1} \in \{1, 2, \dots, n\}$, are continuously differentiable. Their partial derivatives

$$\partial_{i_1} \partial_{i_2} \dots \partial_{i_k} f := \partial_{i_1} \left(\partial_{i_2} \dots \partial_{i_k} f \right) : U \to \mathbb{R}^m$$

are called the *k-th order partial derivatives of* f.

c) If f is k times continuously differentiable, we further put

$$\partial_{i_1, i_2, \dots, i_k} f := \partial_{i_1} \partial_{i_2} \dots \partial_{i_k} f$$

for all $i_1, i_2, \dots, i_k \in \{1, 2, \dots, n\}$ and

$$\partial_i^k f := \underbrace{\partial_i \partial_i \dots \partial_i}_{k \text{ times}} f$$

for every $i \in \{1, 2, \dots, n\}$.

d) Analogously, if f is k times continuously differentiable, we define higher order directional derivatives of the form

$$\partial_{v_1} \partial_{v_2} \dots \partial_{v_k} f : U \to \mathbb{R}^m$$

for all $v_1, \dots, v_k \in \mathbb{R}^n \setminus \{0\}$ in the same way as we defined k-th order partial derivatives. We omit the details.

The following theorem is one of the most important result about partial derivatives that will turn out to be useful both in computations and in theoretical considerations.

Theorem 12.44 (Schwarz)**.** *Let $m, n \in \mathbb{N}$ and let $U \subset \mathbb{R}^n$. Let $f : U \to \mathbb{R}^m$ be twice continuously differentiable and let $a \in U$. Then*

$$\partial_i \partial_j f(a) = \partial_j \partial_i f(a) \qquad \forall i, j \in \{1, 2, \dots, n\}.$$

Proof. We assume w.l.o.g. that

- $m = 1$, since the partial derivatives are computed componentwise,
- $n = 2$, $i = 1$ and $j = 2$, since the general case is proven completely analogously,
- $a = 0 = (0, 0)$, since the line argument is easily transferred to an arbitrary point.

Since U is open, there exists $\delta > 0$ with $B_{2\delta}(0) \subset U$ and one easily derives that $(-\delta,\delta) \times (-\delta,\delta) \subset U$. For a given $y \in (-\delta,\delta)$ let

$$g_y : (-\delta,\delta) \to \mathbb{R}, \qquad g_y(x) = f(x,y) - f(x,0).$$

Since f is continuously differentiable, g_y is differentiable and by definition $g_y'(x) = \partial_1 f(x,y) - \partial_1 f(x,0)$ for any such y. By the mean value theorem from Mathematics 1 (Theorem 4.20) there exists $\xi \in (-\delta,\delta)$ with $|\xi| \leq |x|$, such that

$$g_y(x) - g_y(0) = g_y'(\xi)(x-0) \quad \Leftrightarrow \quad f(x,y) - f(x,0) - f(0,y) + f(0,0) = (\partial_1 f(x,y) - \partial_1 f(x,0))x.$$

We next consider the function $h_\xi : (-\delta,\delta) \to \mathbb{R}$, $h_\xi(y) = \partial_1 f(\xi,y)$ for fixed $\xi \in (-\delta,\delta)$. Since f is twice continuously differentiable, $\partial_1 f$ is differentiable, which yields by construction that h_ξ is differentiable with $h_\xi'(y) = \partial_2\partial_1 f(\xi,y)$. By the mean value theorem applied to h_ξ for any such y there exists $\eta \in (-\delta,\delta)$ with $|\eta| \leq |y|$, such that

$$h_\xi(y) - h_\xi(0) = h_\xi'(\eta)(y-0) \quad \Leftrightarrow \quad \partial_1 f(\xi,y) - \partial_1 f(\xi,0) = \partial_2\partial_1 f(\xi,\eta)y.$$

Inserting this into the above computation shows that for all $x,y \in (-\delta,\delta)$ there are $\xi(x,y), \eta(x,y) \in (-\delta,\delta)$ with $|\xi(x,y)| \leq |x|$ and $|\eta(x,y)| \leq |y|$, such that

$$f(x,y) - f(x,0) - f(0,y) + f(0,0) = \partial_2\partial_1 f(\xi(x,y),\eta(x,y))xy.$$

Analogously, with the roles of the two partial derivatives reversed, we can derive along the same lines that for all $x,y \in (-\delta,\delta)$ there are $\tilde{\xi}(x,y), \tilde{\eta}(x,y) \in (-\delta,\delta)$ with $|\tilde{\xi}(x,y)| \leq |x|$ and $|\tilde{\eta}(x,y)| \leq |y|$, such that

$$f(x,y) - f(x,0) - f(0,y) + f(0,0) = \partial_1\partial_2 f(\tilde{\xi}(x,y),\tilde{\eta}(x,y))xy.$$

If $xy \neq 0$, then this yields

$$\partial_2\partial_1 f(\xi(x,y),\eta(x,y)) = \partial_1\partial_2 f(\tilde{\xi}(x,y),\tilde{\eta}(x,y)). \tag{12.3}$$

One derives from the properties of ξ, $\tilde{\xi}$, η and $\tilde{\eta}$ that

$$\lim_{(x,y)\to(0,0)} \xi(x,y) = \lim_{(x,y)\to(0,0)} \eta(x,y) = \lim_{(x,y)\to(0,0)} \tilde{\xi}(x,y) = \lim_{(x,y)\to(0,0)} \tilde{\eta}(x,y) = (0,0).$$

Thus, since f is twice continuously differentiable, taking limits in (12.3) yields

$$\partial_2\partial_1 f(0,0) = \lim_{(x,y)\to(0,0)} \partial_2\partial_1 f(\xi(x,y),\eta(x,y)) = \lim_{(x,y)\to(0,0)} \partial_1\partial_2 f(\tilde{\xi}(x,y),\tilde{\eta}(x,y)) = x\partial_1\partial_2 f(0,0).$$

□

Schwarz's theorem has a simple consequence for directional derivatives.

Corollary 12.45. *Let $m,n \in \mathbb{N}$ and let $U \subset \mathbb{R}^n$. Let $f : U \to \mathbb{R}^m$ be twice continuously differentiable and let $a \in U$. Then*

$$\partial_v\partial_w f(a) = \partial_w\partial_v f(a) \qquad \forall v,w \in \mathbb{R}^n \setminus \{0\}.$$

Proof. (This proof is not discussed in the lecture videos.)

Let $v = (v_1, \dots, v_n), w = (w_1, \dots, w_n) \in \mathbb{R}^n \setminus \{0\}$. Since f is differentiable, it holds by Theorem 12.15 and Corollary 12.17 that

$$\partial_v f(a) = \sum_{i=1}^{n} \partial_i f(a) v_i, \qquad \partial_w f(a) = \sum_{j=1}^{n} \partial_j f(a) w_j.$$

Since f is twice continuously differentiable, $\partial_i f : U \to \mathbb{R}^m$ is differentiable for each $i \in \{1, 2, \dots, n\}$, so we obtain from the previous computation and from Theorem 12.23 that

$$\begin{aligned} \partial_v \partial_w f(a) &= D(\partial_w f)_a(v) = D\Big(\sum_{j=1}^{n} w_j \partial_j f\Big)_a(v) = \sum_{j=1}^{n} D(\partial_j f)_a(v) w_j \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} \partial_i \partial_j f(a) v_i w_j \overset{\text{Theorem 12.44}}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_j \partial_i f(a) v_i w_j. \end{aligned}$$

Performing the same steps in reverse with the roles of v and w exchanged shows that $\partial_v \partial_w f(a) = \partial_w \partial_v f(a)$. □

12.5 Higher order differentials and Taylor's theorem

To establish a notion of higher differentials, we first discuss another notion from linear algebra.

Definition 12.46. Let K be a field, let V and W be vector spaces over K and let $n \in \mathbb{N}$ with $n \geq 2$.

a) A map

$$f : V^n = \underbrace{V \times V \times \cdots \times V}_{n \text{ times}} \to W$$

is called *n-linear* (*bilinear* for $n = 2$, *trilinear* for $n = 3$, ...) if for all $i \in \{1, 2, \dots, n\}$, $v_1, \dots, v_n, w_i \in K^n$ and all $\lambda, \mu \in K$ it holds that

$$\begin{aligned} &f(v_1, \dots, v_{i-1}, \lambda v_i + \mu w_i, v_{i+1}, \dots, v_n) \\ &= \lambda f(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_n) + \mu f(v_1, \dots, v_{i-1}, w_i, v_{i+1}, \dots, v_n). \end{aligned}$$

We further put $L^n(V, W) := \{f : V^n \to W \mid f \text{ is } n\text{-linear}\}$.

b) For $W = K$, an n-linear map $f : V^n \to K$ is called an *n-linear form on V* (*bilinear form* for $n = 2$, *trilinear form* for $n = 3$,...).

c) A map is called *multilinear* if it is n-linear for some $n \in \mathbb{N}$.

Example 12.47. (1) The map $f : \mathbb{R}^n \to \mathbb{R}$, $f(x_1, \dots, x_n) = x_1 x_2 \dots x_n$, is an n-linear form on $\mathbb{R}$, i.e. $f \in L^n(\mathbb{R}, \mathbb{R})$, for all $n \in \mathbb{N}$.

(2) Let V be a real vector space. Then any inner product $\langle \cdot, \cdot \rangle : V \to \mathbb{R}$ is a bilinear form on V, i.e. $\langle \cdot, \cdot \rangle \in L^2(V, \mathbb{R})$.

(3) The determinant as a map $\det : (\mathbb{R}^n)^n \to \mathbb{R}$ as a map with n vectors as arguments (see Section 9.1) is an n-linear form on $\mathbb{R}^n$, i.e. $\det \in L^n(\mathbb{R}^n, \mathbb{R})$. This follows from property (1) in Definition 9.2.

(4) The vector product $\times : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ that was constructed in Exercise 3 of Sheet 8 is bilinear, i.e. $\cdot \times \cdot \in L^2(\mathbb{R}^3, \mathbb{R}^3)$.

In the same way that we established differentials as linear maps, we will next define the k-th differential as a k-linear map in the following statement.

Theorem/Definition 12.48. *Let $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$ be k times continuously differentiable. Then for each $a \in U$ the map*

$$D^k f_a : (\mathbb{R}^n)^k \to \mathbb{R}^m, \qquad D^k f_a(v_1, \dots, v_k) = \partial_{v_1} \partial_{v_2} \dots \partial_{v_k} f(a),$$

called the k-th differential of f in a, *is k-linear and satisfies*

$$D^k f_a(v_1, \dots, v_k) = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \partial_{i_1, i_2, \dots, i_k} f(a) v_{1 i_1} v_{2 i_2} \dots v_{k i_k} \tag{12.4}$$

for all $v_j = (v_{j1}, v_{j2}, \dots, v_{jn}) \in \mathbb{R}^n$, $j \in \{1, 2, \dots, k\}$. (Here, we use the convention that $\partial_0 g(a) := 0$ for all $g : U \to \mathbb{R}^m$.)

Proof. We want to prove the formula (12.4) by induction over k. The k-linearity of $D^k f_a$ can then be checked from that equation in a straightforward way.

Base case of $k = 1$: This is nothing but Corollary 12.17.

Induction step: Assume as induction hypothesis that

$$D^{k-1} f(a) = \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \partial_{i_2, \dots, i_k} f(a) v_{2 i_2} \dots v_{k i_k}$$

for all $v_j = (v_{j1}, v_{j2}, \dots, v_{jn}) \in \mathbb{R}^n$, $j \in \{2, 3, \dots, k\}$. Since f is k times continuously differentiable, all of the $(k-1)$-st order partial derivatives $\partial_{i_2, \dots, i_k} f(a)$ are continuously differentiable, thus differentiable. Thus, by Theorem 12.15 and Corollary 12.17,

$$\partial_{v_1}(\partial_{i_2, \dots, i_k} f)(a) = D(\partial_{i_2, \dots, i_k} f)_a(v) = \sum_{i_1=1}^{n} \partial_{i_1, i_2, \dots, i_k} f(a) v_{1 i_1} \tag{12.5}$$

for all $v_1 = (v_{11}, v_{12}, \dots, v_{1n}) \in \mathbb{R}^n$. Consequently,

$$\begin{aligned} D^k f_a(v_1, \dots, v_k) &= \partial_{v_1} \partial_{v_2} \dots \partial_{v_k} f(a) = \partial_{v_1}(\partial_{v_2} \dots \partial_{v_k} f)(a) \\ &= \partial_{v_1}(D^{k-1} f_a(v_2, \dots, v_k)) \\ &= \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \partial_v(\partial_{i_2, \dots, i_k} f)(a) v_{2 i_2} \dots v_{k i_k} \\ &\overset{(12.5)}{=} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \partial_{i_1, i_2, \dots, i_k} f(a) v_{1 i_1} v_{2 i_2} \dots v_{k i_k} \end{aligned}$$

for all $v_j = (v_{j1}, v_{j2}, \dots, v_{jn}) \in \mathbb{R}^n$, $j \in \{1, 2, \dots, k\}$. This completes the induction step and thereby shows the claim. □

Remark 12.49. In the case of $n = 2$ and $k = 2$ in the situation of Theorem/Definition 12.48, then the explicit formula for $D^2 f_a$ takes the following relatively concise form for all $a \in U$:

$$D^2 f_a(v, w) = \partial_{1,1} f(a) v_1 w_1 + \partial_{2,1} f(a) v_2 w_1 + \partial_{1,2} f(a) v_1 w_2 + \partial_{2,2} f(a) v_2 w_2$$

for all $v_1 = (v_{11}, v_{12}), v_2 = (v_{21}, v_{22}) \in \mathbb{R}^2$.

Example 12.50. Let $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1^2 + x_2^2$. We have seen in Example 12.9 that

$$\partial_1 f(x_1, x_2) = 2x_1, \quad \partial_2 f(x_1, x_2) = 2x_2.$$

These partial derivatives are again polynomials, hence f is twice continuously differentiable. We compute for all $(x_1, x_2) \in \mathbb{R}^2$ that

$$\partial_{1,1} f(x_1, x_2) = 2, \quad \partial_{2,2} f(x_1, x_2) = 2, \quad \partial_{1,2} f(x_1, x_2) = \partial_{2,1} f(x_1, x_2) = 0.$$

Hence, by Remark 12.49 the second differential of f in $a \in \mathbb{R}^2$ is given by

$$D^2 f_a((v_1, v_2), (w_1, w_2)) = 2v_1 w_1 + 2v_2 w_2 \qquad \forall (v_1, v_2), (w_1, w_2) \in \mathbb{R}^2.$$

We proceed to establish a multivariable version of Taylor's theorem. In the same way as in Mathematics 1, such a theorem tells us about how to approximate a given function by a polynomial - only this time in several variables.

Definition 12.51. Let $k, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be k times continuously differentiable. Given $a \in U$ we call

$$T_k f(\cdot, a) : \mathbb{R}^n \to \mathbb{R}^m, \qquad T_k f(v, a) := f(a) + \sum_{r=1}^{k} \frac{D^r f_a(v, v, \dots, v)}{r!}$$

the k-th Taylor polynomial of f in a.[4]

Remark 12.52. Let $k, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be k times continuously differentiable. Using the explicit formulas for differentials and directional derivatives in terms of partial derivatives, we compute that

$$\begin{aligned} T_2 f(v, a) &= f(a) + Df_a(v) + \frac{D^2 f_a(v, v)}{2} \\ &= f(a) + \sum_{i=1}^{n} \partial_i f(a) v_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_i \partial_j f(a) v_i v_j, \end{aligned}$$

and, more generally,

$$T_k f((v_1, \dots, v_n, a) = f(a) + \sum_{r=1}^{k} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_r=1}^{n} \frac{1}{r!} \partial_{i_1, i_2, \dots, i_r} f(a) v_{i_1} v_{i_2} \dots v_{i_n}.$$

This shows that $T_k f(\cdot, a) : \mathbb{R}^n \to \mathbb{R}$ *is indeed a polynomial of degree at most k* whose coefficients are given by multiples of the partial derivatives of f in a.

We can now use the Lagrange form of Taylor's theorem from Mathematics 1 (Theorem 4.42) for single-variable functions to establish a corresponding result for functions in several variables.

Theorem 12.53 (Taylor, in Lagrange form)**.** *Let $k, n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be $k+1$ times continuously differentiable. Let $a \in U$ and let $h \in \mathbb{R}^n$ be such that $a + th \in U$ for all $t \in [0, 1]$. Then there exists $\theta \in [0, 1]$, such that*

$$f(a + h) = T_k f(h, a) + \frac{1}{(k+1)!} D^{k+1}_{a+\theta h}(h, h, \dots, h).$$

[4]Please note the little difference to the notation Mathematics 1: Here, in the one-dimensional case, $T_k(h, a)$ in the sense of Definition 12.51 would correspond to $T_k(a + h, a)$ in the sense of Definition 4.34. (We apologize for the clash of notations, but this notation is much more convenient in the higher-dimensional setting.)

Proof. Let $g : [0,1] \to \mathbb{R}$, $g(t) = f(a+th)$. Then $g = f \circ \gamma$, where $\gamma : [0,1] \to U$, $\gamma(t) = a + th$. Since γ is obviously k times continuously differentiable, g is k times continuously differentiable. By the Lagrange form of Taylor's theorem for real functions on intervals (Theorem 4.36) for each $x \in [0,1]$ there exists $\theta \in [0,1]$, such that

$$g(x) = \sum_{r=0}^{k} \frac{g^{(r)}(0)}{r!} x^k + \frac{g^{(k+1)}(\theta)}{(k+1)!} x^{k+1},$$

in particular

$$g(1) = g(0) + \sum_{r=1}^{k} \frac{g^{(r)}(0)}{r!} + \frac{g^{(k+1)}(\theta)}{(k+1)!} \quad \Leftrightarrow \quad f(a+h) = f(a) + \sum_{r=1}^{k} \frac{g^{(r)}(0)}{r!} + \frac{g^{(k+1)}(\theta)}{(k+1)!}. \tag{12.6}$$

We need to compute the derivatives of g in 0.

Claim. $g^{(r)}(t) = D^r f_{\gamma(t)}(h,h,\dots,h)$ for all $r \in \{1,2,\dots,k\}$.

Proof of claim. We want to show this by induction over r. For the base case of $r = 1$, we compute using Corollary 12.29 that

$$g'(t) = (f \circ \gamma)'(0) = Df_{\gamma(t)}(\gamma'(t)) = Df_{\gamma(t)}(h) = \sum_{i=1}^{n} \partial_i f(\gamma(t)) h_i,$$

where $h = (h_1, \dots, h_n)$. In particular, $g'(0) = Df_a(h)$, which shows the base case.
In the induction step, we assume as induction hypothesis that $g^{(r)}(t) = D^r f_{\gamma(t)}(h,h,\dots,h)$ has been shown for some $r \in \{1,2,\dots,k-1\}$. By Theorem/Definition 12.48, it then holds that

$$g^{(r)}(t) = \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} \partial_{i_1,\dots,i_r} f(\gamma(t)) h_{i_1} \dots h_{i_r}$$

Since $r < k$ and f is k times continuously differentiable, $g^{(r)}$ is again differentiable. From the usual rules for derivatives and from Corollary 12.29 and Corollary 12.17 we derive:

$$\begin{aligned} g^{(r+1)}(t) &= \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} (\partial_{i_1,\dots,i_r} f \circ \gamma)'(t) h_{i_1} \dots h_{i_r} \\ &= \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} D(\partial_{i_1,\dots,i_r} f)_{\gamma(t)}(\gamma'(t)) h_{i_1} \dots h_{i_r} \\ &= \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} \sum_{i_{r+1}=1}^{n} \partial_{i_{r+1},i_1,\dots,i_r} f(\gamma(t)) h_{i_{r+1}} h_{i_1} \dots h_{i_r}. \end{aligned}$$

One derives from iteratively using Schwarz's theorem (Theorem 12.44) that

$$\partial_{i_{r+1},i_1,\dots,i_r} f = \partial_{i_1,i_{r+1},i_2,\dots,i_r} f = \cdots = \partial_{i_1,\dots,i_r,i_{r+1}} f \qquad \forall i_1,\dots,i_r,i_{r+1} \in \{1,2,\dots,n\}.$$

Hence,

$$g^{(r+1)}(t) = \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} \sum_{i_{r+1}=1}^{n} \partial_{i_1,\dots,i_r,i_{r+1}} f(\gamma(t)) h_{i_1} \dots h_{i_r} h_{i_{r+1}} = D^{r+1} f_{\gamma(t)}(h,h,\dots,h),$$

where we used Theorem/Definition 12.48 for the last equality. This completes the induction step. ■

It follows from the above claim that $g^{(r)}(0) = D^r f_a(h,h,\dots,h)$ for all $r \in \{1,2,\dots,k\}$. Moreover,

$$g^{(k+1)}(\theta) = D^{k+1} f_{\gamma(\theta)}(h,h,\dots,h) = D^{k+1} f_{a+\theta h}(h,h,\dots,h).$$

Inserting these computations into (12.6) shows the claim. □

To discuss approximation properties of Taylor polynomials, we also establish a Péano form of Taylor's theorem, corresponding to Theorem 4.36 from Mathematics 1.

Theorem 12.54 (Taylor, in Péano form)**.** *Let $k,n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ be open, $a \in U$ and $f : U \to \mathbb{R}$ be k times continuously differentiable. Then, with $U' := \{h \in \mathbb{R}^n \mid a+th \in U\ \forall t \in [0,1]\}$, there exists $\varphi_k : U' \to \mathbb{R}$, so that*

$$f(a+h) = T_k f(h,a) + \varphi_k(h) \qquad \forall h \in U',$$

and so that

$$\lim_{h\to 0} \frac{\varphi_k(h)}{\|h\|^k} = 0.$$

Proof. (This proof is not discussed in the lecture videos.)

We define $\varphi_k : U' \to \mathbb{R}^m$, $\varphi_k := f(a+h) - T_k f(h,a)$. We need to show that φ_k has the necessary limiting property from the claim. Since U is open, there exists $r > 0$, such that $B_r(a) \subset U$ and to study the limit, we can assume w.l.o.g. that $U = B_r(a)$. Then $U' = B_r(0)$, so U' is convex. Let $\varepsilon > 0$. Since f is k times continuously differentiable, all of its k-th order partial derivatives are continuous. Thus for all $i_1,\dots,i_k \in \{1,2,\dots,n\}$ there exists $\delta_{i_1,\dots,i_k} > 0$, such that

$$\left|\partial_{i_1,\dots,i_k} f(y) - \partial_{i_1,\dots,i_k} f(a)\right| < \frac{\varepsilon}{kn} k! \qquad \forall y \in B_{\delta_{i_1,\dots,i_k}}(a).$$

If we put $\delta := \min\{\delta_{i_1,\dots,i_k} \mid i_1,\dots,i_k \in \{1,2,\dots,n\}\} > 0$, then

$$\frac{1}{k!} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \left|\partial_{i_1,\dots,i_k} f(y) - \partial_{i_1,\dots,i_k} f(a)\right| < \varepsilon \qquad \forall y \in B_\delta(a). \tag{12.7}$$

To make use of this estimate, we derive from Theorem/Definition 12.48 that for all $h \in U'$

$$\begin{aligned}
\left|\frac{D^k f_y(h,h,\dots,h)}{k!} - \frac{D^k f_a(h,h,\dots,h)}{k!}\right| &= \left|\sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \left(\partial_{i_1,i_2,\dots,i_k} f(y) - \partial_{i_1,i_2,\dots,i_k} f(a)\right) h_{i_1} h_{i_2} \dots h_{i_k}\right| \\
&\le \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \left|\partial_{i_1,i_2,\dots,i_k} f(y) - \partial_{i_1,i_2,\dots,i_k} f(a)\right| |h_{i_1}||h_{i_2}| \dots |h_{i_k}| \\
&\le \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \left|\partial_{i_1,i_2,\dots,i_k} f(y) - \partial_{i_1,i_2,\dots,i_k} f(a)\right| \|h\|^k.
\end{aligned}$$

Hence, it follows from (12.7) that

$$\frac{1}{k!}\left|\frac{D^k f_y(h,h,\dots,h)}{k!} - \frac{D^k f_a(h,h,\dots,h)}{k!}\right| < \varepsilon \|h\|^k \qquad \forall y \in B_\delta(a), h \in U'. \tag{12.8}$$

Then by Theorem 12.53 for each $h \in U'$ there exists $\theta \in [0,1]$ with

$$f(a+h) = T_{k-1}f(h,a) + \frac{D^k f_{a+\theta h}(h,h,\dots,h)}{k!},$$

which yields

$$f(a+h) = T_k f(h,a) + \frac{D^k f_{a+\theta h}(h,h,\dots,h)}{k!} - \frac{D^k f_a(h,h,\dots,h)}{k!}$$

and thus

$$|\varphi_k(h)| = \frac{1}{k!}\left|D^k f_{a+\theta h}(h,h,\dots,h) - D^k f_a(h,h,\dots,h)\right|.$$

Now if $h \in B_\delta(0)$, then $a + \theta h \in B_\delta(a)$, so it follows from (12.8) that

$$|\varphi_k(h)| < \varepsilon \|h\|^k \qquad \forall h \in B_\delta(0)$$

and thus

$$\lim_{h\to 0} \frac{|\varphi_k(h)|}{\|h\|^k} \leq \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that $\lim_{h\to 0} \frac{|\varphi_k(h)|}{\|h\|^k} = 0$ and thus $\lim_{h\to 0} \frac{\varphi_k(h)}{\|h\|^k} = 0$ as we wanted to show. □

Remark 12.55. We can interpret Theorem 12.54 in such a way that, in analogy with the results from Mathematics 1 for functions on intervals, the Taylor polynomial $T_k f(\cdot, a)$ is *the best approximation of f near a by a polynomial of degree at most k.*

To conclude this section, we will compute the second Taylor polynomial and compute its approximation properties for a concrete example in excessive detail.

Example 12.56. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = \sin(x^2y + z)$. We want to compute the second Taylor polynomial of f in $(1,0,0)$.

$$\partial_1 f(x,y,z) = 2xy\cos(x^2y+z), \quad \partial_2 f(x,y,z) = x^2\cos(x^2y+z), \quad \partial_3 f(x,y,z) = \cos(x^2y+z).$$

Using the computational rules for partial derivatives and Schwarz's theorem, one computes that

$$\begin{aligned}
\partial_{1,1} f(x,y,z) &= 2y\cos(x^2y+z) - 4x^2y^2\sin(x^2y+z),\\
\partial_{1,2} f(x,y,z) &= \partial_{2,1} f(x,y,z) = 2x\cos(x^2y+z) - 2x^3y\sin(x^2y+z),\\
\partial_{1,3} f(x,y,z) &= \partial_{3,1} f(z,y,z) = -2xy\sin(x^2y+z),\\
\partial_{2,2} f(x,y,z) &= -x^4\sin(x^2y+z),\\
\partial_{2,3} f(x,y,z) &= \partial_{3,2} f(x,y,z) = -x^2\sin(x^2y+z),\\
\partial_{3,3} f(x,y,z) &= -\sin(x^2y+z).
\end{aligned}$$

Using the formula from Remark 12.52 we compute for $v = (v_1, v_2, v_3) \in \mathbb{R}^3$ that

$$T_2 f(v, (1,0,0)) = f(1,0,0) + \sum_{i=1}^{3} \partial_i f(1,0,0) v_i + \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \partial_{i,j} f(1,0,0) v_i v_j$$

$$= f(1,0,0) + \partial_1 f(1,0,0) v_1 + \partial_2 f(1,0,0) v_2 + \partial_3 f(1,0,0) v_3 + \frac{1}{2} \Big(\partial_{1,1} f(1,0,0) v_1^2 + 2\partial_{1,2} f(1,0,0) v_1 v_2$$
$$+ 2\partial_{1,3}(1,0,0) v_1 v_3 + \partial_{2,2} f(1,0,0) v_2^2 + 2\partial_{2,3} f(1,0,0) v_2 v_3 + \partial_{3,3} f(1,0,0) v_3^2 \Big)$$

$$= 0 + 0 \cdot \cos(0) v_1 + 1 \cdot \cos(0) v_2 + \cos(0) v_3 + \frac{1}{2} \Big((0 \cdot \cos(0) - 0 \cdot \sin(0)) v_1^2$$

$$+ 2(2\cos(0) - 0 \cdot \sin(0)) v_1 v_2 + 0 \cdot \sin(0) v_1 v_3 - 1 \cdot \sin(0) v_2^2 + 2(-\sin(0)) v_2 v_3 - \sin(0) v_3^2 \Big)$$

$$= v_2 + v_3 + 2 v_1 v_2.$$

Thus, $T_2 f(v, (1,0,0)) = v_2 + v_3 + 2v_1 v_2$ is the best approximation of f near $(1,0,0)$ by a polynomial of degree two.

(The rest of this example is not discussed in the lecture videos.)

Let's take a look at how much this polynomial deviates from f near $(1,0,0)$. Put

$$\varphi(v) := f((1,0,0) + v) - T_2 f(v, (1,0,0)$$

for $v = (v_1, v_2, v_3) \in \mathbb{R}^3$. By the previous computations,

$$\varphi(v) = \sin((v_1 + 1)^2 v_2 + v_3) - v_2 - v_3 + 2 v_1 v_2.$$

We have seen in Theorem 3.80 of Mathematics 1 that there exists $r : \mathbb{R} \to \mathbb{R}$ with $|r(x)| \leq \frac{|x|^3}{6}$ for $x \in [-4, 4]$, such that $\sin(x) = x + r(x)$. In terms of this map, we compute that

$$\begin{aligned} \varphi(v) &= (v_1 + 1)^2 v_2 + v_3 + r((v_1 + 1)^2 v_2 + v_3) - v_2 - v_3 - 2 v_1 v_2 \\ &= (v_1^2 + 2v_1 + 1) v_2 + v_3 - v_2 - v_3 - 2 v_1 v_2 + r((v_1 + 1)^2 v_2 + v_3) \\ &= v_1^2 v_2 + r((v_1 + 1)^2 v_2 + v_3). \end{aligned}$$

Thus, if $|(v_1 + 1)^2 v_2 + v_3| \leq 4$, then

$$|\varphi(v)| \leq |v_1|^2 |v_2| + |r(v)| \leq |v_1|^2 + \frac{1}{6} |(v_1 + 1)^2 v_2 + v_3|^3. \tag{12.9}$$

We compute using that $|v_i| \leq \|v\|$ for $i \in \{1, 2, 3\}$ that

$$\begin{aligned} |(v_1 + 1)^2 v_2 + v_3| &\leq (v_1 + 1)^2 |v_2| + |v_3| \leq (v_1^2 + 2|v_1| + 1) |v_2| + |v_3| \\ &\leq (\|v\|^2 + 2\|v\| + 1)\|v\| + \|v\| = (\|v\|^2 + 2\|v\| + 2)\|v\| \end{aligned}$$

Inserting this into (12.9) yields

$$|\varphi(v)| \leq \|v\|^3 + \frac{1}{6} (\|v\|^2 + 2\|v\| + 2)^3 \|v\|^3.$$

Thus, if $v \in B_{1/10}(0)$, then

$$\begin{aligned} |f((1,0,0) + v) - T_2 f(v, (1,0,0))| = |\varphi(v)| &\leq \frac{1}{1000} + \frac{1}{6} \left(\frac{1}{100} + \frac{1}{10} + 2 \right)^3 \frac{1}{1000} \\ &\leq \frac{1}{1000} + \frac{1}{6000} \left(\frac{5}{4} \right)^3 = \frac{1}{1000} + \frac{1}{3072} \approx 0,0013. \end{aligned}$$

Hence, in a ball of radius $\frac{1}{10}$ around $(1,0,0)$, the degree-two polynomial $T_2f(v,(1,0,0)$ approximates f up to a little more than $\frac{1}{1000}$. Not bad!

12.6 The second differential and local extrema of a function

We have learnt in Mathematics 1 that the local extrema of a twice differentiable function on an interval can be determined by studying the derivatives of f. The necessary condition for local maximum/minimum points was that $f'(x_0) = 0$. We have generalized this observation to functions in several variables in Theorem 12.36, and observed that the gradient of a function must vanish at local maximum/minimum points.
In the single-variable case, we have further seen that the question if a point with vanishing derivative is a local maximum point, a local minimum point or neither of the two is decided by the first non-vanishing higher derivative in x_0. In this section, we want to generalize this method to functions in several variables, which requires some preparations.
Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be twice continuously differentiable. Then its second differential $D^2f_a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ in $a \in U$ is a bilinear form on $\mathbb{R}^n$. To decide from this bilinear form if a point is a local maximum point, a local minimum point or neither of the two, we first need to discuss certain results from the linear algebra of bilinear forms.

12.6.1 Bilinear forms

Let K be a field and V be a vector space over K. We recall from Definition 12.46 that a bilinear form on V is a map $s : V \times V \to K$, for which

$$s(\lambda_1 v_1 + \lambda_2 v_2, w) = \lambda_1 s(v_1, w) + \lambda_2 s(v_2, w) \qquad \forall \lambda_1, \lambda_2 \in K,\ v_1, v_2, w \in V$$

Definition 12.57. Let V be a real vector space and let $s : V \times V \to \mathbb{R}$ be a bilinear form.

a) s is called *symmetric* if $s(v,w) = s(w,v)$ for all $v, w \in K$.

b) s is called

- *positive definite* if $s(v,v) > 0$ for all $v \in V$,
- *negative definite* if $s(v,v) < 0$ for all $v \in V$,
- *indefinite* if there exist $v_1, v_2 \in V$ with $s(v_1, v_1) > 0$ and $s(v_2, v_2) < 0$.

Remark 12.58. At this point, we can rephrase the definition one of our most important notions from linear algebra: An inner product on a real vector space V is the same as a symmetric and positive definite bilinear form on V.

Example 12.59. (1) Consider the maps $s_1, s_2, s_3, s_4 : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, where

$$\begin{aligned} s_1((x_1,y_1),(x_2,y_2)) &= x_1x_2 + y_1y_2, & s_2((x_1,y_1),(x_2,y_2)) &= -x_1x_2 - y_1y_2, \\ s_3((x_1,y_1),(x_2,y_2)) &= x_1x_2 - y_1y_2, & s_4((x_1,y_1),(x_2,y_2)) &= x_1x_2. \end{aligned}$$

One checks that all four maps are symmetric bilinear forms. Moreover, it is easy to see that s_1 is positive definite, s_2 is negative definite and s_3 is indefinite, while s_4 is neither of the three, since $s_4(x,x) \geq 0$ for all $x \in \mathbb{R}^2$ with $s_4((0,y),(0,y)) = 0$ for all $y \in \mathbb{R}$.

(2) The map $s : \mathbb{R}^2 \to \mathbb{R}^2 \to \mathbb{R}$, $s((x_1, y_1), (x_2, y_2)) = x_1 y_2 - x_2 y_1$, is a bilinear form which is not symmetric, since $s(x, y) = -s(y, x)$ for all $x, y \in \mathbb{R}^2$. Moreover, $s(x, x) = 0$ for all $x \in \mathbb{R}^2$, so s has none of the three definiteness properties.[5]

Given a Euclidean vector space, it turns out that any bilinear form is related to the inner product by an endomorphism, as we shall make precise in the following result.

Theorem 12.60. *Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional Euclidean vector space, let $n = \dim V$ and let $s : V \times V \to \mathbb{R}$ be a bilinear form. Then:*

a) There exists an endomorphism $f : V \to V$, such that

$$s(x, y) = \langle x, f(y) \rangle \qquad \forall x, y \in V.$$

b) s is symmetric if and only if f is self-adjoint.

Proof. a) Let $y \in V$. Since s is bilinear, the map $s(\cdot, y) : V \to \mathbb{R}$ is linear, so by the Riesz Representation Theorem, there is a unique $v_y \in V$, such that $s(x, y) = \langle x, v_y \rangle$ for all $x \in V$. Since such a v_y exists for each $y \in V$, we can define a map $f : V \to V$ by $f(y) = v_y$. Then $s(x, y) = \langle x, f(y) \rangle$ for all $x, y \in V$ and it remains to show that this map is an endomorphism. For all $x, y_1, y_2 \in V$ we compute that

$$\langle x, f(y_1) + f(y_2) \rangle = \langle x, v_{y_1} + v_{y_2} \rangle = \langle x, v_{y_1} \rangle + \langle x, v_{y_2} \rangle = s(x, y_1) + s(x, y_2) = s(x, y_1 + y_2)$$

using the bilinearity of s and of the inner product. By the uniqueness part of the Riesz representation theorem, this shows that $f(y_1) + f(y_2) = v_{y_1 + y_2} = f(y_1 + y_2)$. Analogously, one uses the bilinearity properties to show that $f(\lambda y) = \lambda f(y)$ for all $\lambda \in \mathbb{R}$ and $y \in V$. Hence, f is an endomorphism.

b) By the symmetry of inner products and the definition of f, we compute that

$$s(x, y) = s(y, x) \ \ \forall x, y \in V \quad \Leftrightarrow \quad \langle x, f(y) \rangle = \langle y, f(x) \rangle = \langle f(x), y \rangle \ \ \forall x, y \in V.$$

This is exactly what we wanted to show.

□

In the following, we will only require the case of $V = \mathbb{R}^n$ with the standard inner product. Since every endomorphism of $\mathbb{R}^n$ is induced by a real $n \times n$-matrix, in that case Theorem 12.60 takes the following form.

Corollary 12.61. *Let $n \in \mathbb{N}$, let $\langle \cdot, \cdot \rangle$ be the standard inner product on $\mathbb{R}^n$ and let $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a bilinear form. Then:*

a) There exists $A \in M(n \times n, \mathbb{R})$, such that

$$s(x, y) = \langle x, Ay \rangle \quad \forall x, y \in \mathbb{R}^n.$$

b) s is symmetric if and only if A is symmetric.

c) The entries of A are given by $A = (a_{ij})$, where $a_{ij} := s(e_i, e_j)$ for all $i, j \in \{1, 2, \ldots, n\}$.

[5] This bilinear form is the simplest example of what is called a *symplectic form*. This notion is important in and grew out of the formalism of Hamiltonian mechanics.

Proof. a) This follows from Theorem 12.60.a) and the fact that every endomorphism of $\mathbb{R}^n$ is induced by a real $n \times n$-matrix.

b) This follows from combining Theorems 12.60.b) and 10.53.

c) We compute for all $i, j \in \{1, 2, \ldots, n\}$ that

$$s(e_i, e_j) = \langle e_i, Ae_j \rangle = \langle (0, \ldots, 0, 1, 0, \ldots, 0), (a_{1j}, a_{2j}, \ldots, a_{nj}) \rangle = a_{ij}.$$

□

Thus, bilinear forms on $\mathbb{R}^n$ are represented by matrix and the definiteness properties of bilinear forms can be "converted" into the following definiteness properties of matrices.

Definition 12.62. Let $n \in \mathbb{N}$ and $A \in M(n \times n, \mathbb{R})$.

a) A is *positive definite* if $\langle x, Ax \rangle > 0$ for all $x \in \mathbb{R}^n$.

b) A is *negative definite* if $\langle x, Ax \rangle < 0$ for all $x \in \mathbb{R}^n$.

c) A is *indefinite* if there are $x, y \in \mathbb{R}^n$ with $\langle x, Ax \rangle > 0$ and $\langle y, Ay \rangle < 0$.

Remark 12.63. Note that by definition, $A \in M(n \times n, \mathbb{R})$ is positive definite/negative definite/indefinite if and only if the bilinear form $s(x, y) = \langle x, Ay \rangle$ on $\mathbb{R}^n$ has the same property.

For symmetric matrices, our knowledge from linear algebra shows that definiteness properties of matrices can be expressed in terms of eigenvalues as well.

Theorem 12.64. *Let $n \in \mathbb{N}$ and let $A \in M(n \times n, \mathbb{R})$ be symmetric. Then:*

a) A is positive definite if and only if all eigenvalues of A are positive.

b) A is negative definite if and only if all eigenvalues of A are negative.

c) A is indefinite if and only if A has a positive and a negative eigenvalue.

Proof. a) By the spectral theorem, A is diagonalizable and $\mathbb{R}^n$ has an orthonormal eigenbasis $(b_1, \ldots, b_n)$ for A. Let λ_i be the eigenvalue corresponding to b_i for each $i \in \{1, 2, \ldots, n\}$. Then for all $\mu_1, \ldots, \mu_n \in \mathbb{R}$ we compute with $x = \sum_{i=1}^n \mu_i b_i$ that

$$\begin{aligned}\langle x, Ax \rangle &= \left\langle \sum_{i=1}^n \mu_i b_i, A\left(\sum_{j=1}^n \mu_j b_j\right) \right\rangle = \sum_{i=1}^n \mu_i \left\langle b_i, \sum_{j=1}^n \mu_j A(b_j) \right\rangle = \sum_{i=1}^n \mu_i \left\langle b_i, \sum_{j=1}^n \mu_j \lambda_j b_j \right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_j \mu_i \mu_j \langle b_i, b_j \rangle = \sum_{i=1}^n \lambda_i \mu_i^2.\end{aligned}$$

Hence,

$$\begin{aligned}\langle x, Ax \rangle > 0 \; \forall x \in \mathbb{R}^n \setminus \{0\} \quad &\Leftrightarrow \quad \sum_{i=1}^n \lambda_i \mu_i^2 > 0 \quad \forall (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n \setminus \{0\} \\ &\Leftrightarrow \quad \lambda_i > 0 \quad \forall i \in \{1, 2, \ldots, n\},\end{aligned}$$

which shows the claim.

b) This is shown in complete analogy with a) or by applying a) to $-A$.

c) "⇒": This is obvious.

"⇐": As in the proof of a), there is an orthonormal eigenbasis $(b_1, \dots, b_n)$ of $\mathbb{R}^n$ with associated eigenvalues $\lambda_1, \dots, \lambda_n$. Then for all $x = \sum_{i=1}^n \mu_i b_i \in \mathbb{R}^n$, we have shown in a) that $\langle x, Ax \rangle = \sum_{i=1}^n \lambda_i \mu_i^2$. Thus, if there exists $x \in \mathbb{R}^n$ with $s(x,x) > 0$, there must be some $i \in \{1, 2, \dots, n\}$ with $\lambda_i > 0$. Analogously, if there exists some $y \in \mathbb{R}^n$ with $s(y,y) < 0$, then there exists an i with $\lambda_i < 0$, which shows the claim.

□

In the case of $n = 2$, there are simple explicit criteria to check the positivity and negativity of the eigenvalues of a real 2×2-matrix.

Corollary 12.65. *Let* $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in M(2 \times 2, \mathbb{R})$ *be symmetric. Then*

a) A is positive definite if and only if $\det A > 0$ *and* $a + c > 0$.

b) A is negative definite if and only if $\det A > 0$ *and* $a + c < 0$.

c) A is indefinite if and only if $\det A < 0$.

Proof. (This proof is not discussed in the lecture videos.)

We compute the characteristic polynomial of A as

$$\chi_A(t) = \det \begin{pmatrix} a - t & b \\ b & c - t \end{pmatrix} = (a - t)(c - t) - b^2 = t^2 - (a + c)t + ac - b^2.$$

The usual solution methods for quadratic equations show that the eigenvalues of A, i.e. the solutions of $\chi_A(t) = 0$ are given by

$$t = \frac{a + c}{2} \pm \sqrt{\frac{1}{4}(a + c)^2 + b^2 - ac} = \frac{a + c}{2} \pm \sqrt{\frac{1}{4}(a - c)^2 + b^2}.$$

Hence, t has the two eigenvalues

$$\lambda_1 = \frac{a + c}{2} + \sqrt{\frac{1}{4}(a - c)^2 + b^2} \quad \vee \quad \lambda_2 = \frac{a + c}{2} - \sqrt{\frac{1}{4}(a - c)^2 + b^2}.$$

a) "⇒": If A is positive definite, then $\lambda_1 > 0$ and $\lambda_2 > 0$ by Theorem 12.64.a). Since A is then similar to $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, it follows from Corollary 9.15.b) that

$$\det A = \det \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \lambda_1 \lambda_2 > 0.$$

Since $\frac{a+c}{2} > \lambda_2 > 0$ it further follows that $a + c > 0$.

"⇐" If $a + c > 0$, then $\lambda_1 > 0$. Moreover, if $\det A = ac - b^2 > 0$, i.e. $b^2 < ac$, then

$$\lambda_2 > \frac{a + c}{2} - \sqrt{\frac{1}{4}(a - c)^2 + ac} = \frac{a + c}{2} - \sqrt{\frac{1}{4}(a + c)^2} = 0.$$

Hence, A is positive definite by Theorem 12.64.a).

b) This is shown in analogy with part a).

c) "⇒": By Theorem 12.64.c), it holds that $\lambda_1 > 0$ and $\lambda_2 < 0$, such that $\det A = \lambda_1\lambda_2 < 0$. "⇐": If $\det A = ac - b^2 < 0$, i.e. $b^2 > ac$, then

$$\lambda_1 > \frac{a+c}{2} + \sqrt{\frac{1}{4}(a-c)^2 + ac} = \frac{a+c}{2} + \sqrt{\frac{1}{4}(a+c)^2} = \frac{a+c}{2} + \left|\frac{a+c}{2}\right| \geq 0$$

and

$$\begin{aligned}\lambda_2 &= \frac{a+c}{2} - \sqrt{\frac{1}{4}(a^2 - 2ac + c^2) + b^2} = \frac{a+c}{2} - \sqrt{\frac{1}{4}(a^2 + 2ac + c^2) + b^2 - ac} \\ &= \frac{a+c}{2} - \sqrt{\left(\frac{a+c}{2}\right)^2 + b^2 - ac} < \frac{a+c}{2} - \sqrt{\left(\frac{a+c}{2}\right)^2} \leq 0.\end{aligned}$$

Hence A is indefinite by Theorem 12.64.c).

□

12.6.2 Local extrema of functions in several variables

We want to apply the results and notions from the previous subsection to the second differential of a function. We summarize the foundation for this approach in the following statement.

Corollary 12.66. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be twice continuously differentiable. Then for each $a \in U$ the second differential $D^2 f_a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a symmetric bilinear form on $\mathbb{R}^n$.*

Proof. By Theorem/Definition 12.48, $D^2 f_a$ is a bilinear form on $\mathbb{R}^n$. By Corollary 12.45,

$$D^2 f_a(v, w) = \partial_v \partial_w f(a) = \partial_w \partial_v f(a) = D^2 f_a(w, v) \qquad \forall v, w \in \mathbb{R}^n,$$

hence $D^2 f_a$ is symmetric. □

We introduce one more bit of terminology before taking our way towards local extrema of functions in several variables..

Definition 12.67. Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be partially differentiable. We call $a \in U$ a *critical point of f* if $\nabla f(a) = 0$. We further put

$$\operatorname{Crit} f = \{a \in U \mid \nabla f(a) = 0\}.$$

Theorem 12.68. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be twice continuously differentiable. Let $a \in U$ be a critical point of f.*

a) If $D^2 f_a$ is positive definite, then f has a local minimum in a.

b) If $D^2 f_a$ is negative definite, then f has a local maximum in a.

c) If $D^2 f_a$ is indefinite, then f has neither a local maximum nor a local minimum in a.

Proof. Let $a \in U$ with $\nabla f(a) = 0$. Then $Df_a = 0$ and by Theorem 12.54, i.e. the Péano form of Taylor's theorem, it holds that

$$f(a+h) = f(a) + \frac{1}{2}D^2 f_a(h,h) + \varphi(h), \tag{12.10}$$

where, with $U' = \{h \in \mathbb{R}^n \mid a+h \in U\}$, $\varphi : U' \to \mathbb{R}$ satisfies $\lim_{h\to 0} \frac{\varphi(h)}{\|h\|^2} = 0$.

a) Let $S := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. One checks that S is closed and bounded in $\mathbb{R}^n$, hence S is compact by the Heine-Borel theorem. Since f is twice continuously differentiable, the map $g : S \to \mathbb{R}$, $g(v) = D^2 f_a(v,v)$, is continuous, hence attains its minimum $m \in \mathbb{R}$. Since $D^2 f_a$ is positive definite, it follows that $m > 0$. Since for all $h \in \mathbb{R}^n \setminus \{0\}$ it holds that $\frac{1}{\|h\|}h \in S$, we compute from the bilinearity of $D^2 f_a$ that for all $h \in \mathbb{R}^n \setminus \{0\}$,

$$D^2 f_a(h,h) = \|h\|^2 D^2 f_a(\tfrac{1}{\|h\|}h, \tfrac{1}{\|h\|}h) = \|h\|^2 g(\tfrac{1}{\|h\|}h) \geq m\|h\|^2.$$

Thus,

$$D^2 f_a(h,h) \geq m\|h\|^2 \qquad \forall h \in \mathbb{R}^n,$$

since the statement is obvious for $h = 0$. Since $\lim_{h\to 0} \frac{\varphi(h)}{\|h\|^2} = 0$, there exists an $r > 0$ such that $B_r(0) \subset U'$ and such that

$$|\varphi_2(x)| \leq \frac{1}{4}m\|h\|^2 \qquad \forall h \in B_r(0).$$

Inserting the obtained inequalities into (12.10) yields

$$f(a+h) \geq f(a) + \frac{1}{2}m\|h\|^2 - \frac{1}{4}m\|h\|^2 = f(a) + \frac{1}{4}m\|h\|^2 \geq f(a) \qquad \forall h \in B_r(0).$$

Hence, a is a local minimum point of f.

b) This follows by applying a) to $g := -f$, since one derives from $D^2 g_a = -D^2 f_a$ that $D^2 g_a$ is positive definite if and only if $D^2 f_a$ is negative definite.

c) Let $v_1 = (v_{11}, v_{12}, \ldots, v_{1n}), v_2 = (v_{21}, v_{22}, \ldots, v_{2n}) \in \mathbb{R}^n$ be given so that $D^2 f_a(v_1, v_1) > 0$ and $D^2 f_a(v_2, v_2) < 0$. For sufficiently small $\varepsilon > 0$ we consider the curves $\gamma_1, \gamma_2 : (-\varepsilon, \varepsilon) \to U$, $\gamma_i(t) = a + tv_i$, and the functions $g_1, g_2 : (-\varepsilon, \varepsilon) \to \mathbb{R}$, $g_i = f \circ \gamma_i$. Then $g_1(0) = g_2(0) = a$. Since f is twice continuously differentiable, it follows that g_1 and g_2 are twice continuously differentiable as well. We compute from Corollary 12.29 that

$$g_i'(t) = \langle \nabla f(\gamma_i(t)), \gamma_i'(t) \rangle = \langle \nabla f(\gamma_i(t)), v_i \rangle = \sum_{j=1}^{n} \partial_j f(\gamma_i(t)) v_{ij} \qquad \forall t \in (-\varepsilon, \varepsilon),\ i \in \{1,2\}.$$

In particular, $g_i'(0) = \langle \nabla f(a), v_i \rangle = 0$ for $i \in \{1,2\}$. We further compute in the same way that

$$\begin{aligned} g_i''(t) &= \sum_{j=1}^{n} \langle \nabla(\partial_j f)(\gamma_i(t)), \gamma_i'(t) \rangle v_{ij} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \partial_j \partial_k f(\gamma_i(t)) v_{ij} v_{ik} = D^2 f_{\gamma_i(t)}(v_i, v_i) \qquad \forall t \in (-\varepsilon, \varepsilon),\ i \in \{1,2\}, \end{aligned}$$

where we used the formula from Theorem/Definition 12.48. In particular, $g_1''(0) = D^2 f_a(v_1, v_1) > 0$ and $g_2''(0) = D^2 f_a(v_2, v_2) < 0$. Thus, by the results of Mathematics 1, g_1 has a local minimum in a while g_2 has a local maximum in f. One further observes that for sufficiently small $t \in (-\varepsilon, \varepsilon)$ with $t \neq 0$, it holds that $f(a + tv_1) = g_1(t) > g_1(0) = f(a)$ and $f(a + tv_2) = g_2(t) < g_2(0) = f(a)$. Hence, f is neither a local maximum point nor a local minimum point of f.

□

As we have seen in the previous section, we can study the definiteness of bilinear forms by studying certain matrices instead. For second differentials, the corresponding matrices can be computed straight from the partial derivatives of the function.

Theorem/Definition 12.69. *Let $n \in \mathbb{N}$, $U \subset \mathbb{R}^n$ open, $a \in U$ and let $f : U \to \mathbb{R}$ be twice continuously differentiable. Then*

$$D^2 f_a(v, w) = \left\langle v, H_f(a) w \right\rangle \qquad \forall v, w \in \mathbb{R}^n,$$

where

$$H_f(a) \in M(n \times n, \mathbb{R}), \qquad H_f(a) = (\partial_i \partial_j f(a)) = \begin{pmatrix} \partial_1\partial_1 f(a) & \partial_1\partial_2 f(a) & \dots & \partial_1\partial_n f(a) \\ \partial_2\partial_1 f(a) & \partial_2\partial_2 f(a) & \dots & \partial_2\partial_n f(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_n\partial_1 f(a) & \partial_n\partial_2 f(a) & \dots & \partial_n\partial_n f(a) \end{pmatrix}.$$

$H_f(a)$ is called the Hessian[6] matrix *or simply the* Hessian of f in a.

Proof. This follows in a straightforward way from Corollary 12.61, since $D^2 f_a(e_i, e_j) = \partial_i \partial_j f(a)$ for all $i \in \{1, 2, \dots, n\}$ by definition of the second differential. □

We can thus use the notions from Subsection 12.6.1 to rephrase the assertions of Theorem 12.68 in terms of Hessian matrices.

Corollary 12.70. *Let $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be twice continuously differentiable. Let $a \in U$ with $\nabla f(a) = 0$.*

a) If $H_f(a)$ is positive definite, then f has a local minimum in a.

b) If $H_f(a)$ is negative definite, then f has a local maximum in a.

c) If $H_f(a)$ is indefinite, then f has neither a local maximum nor a local minimum in a.

Proof. This follows from combining Theorem 12.68 with Theorem/Definition 12.69. □

Example 12.71. (1) Consider the functions

$$f_1, f_2, f_3 : \mathbb{R}^2 \to \mathbb{R}, \quad f_1(x, y) = x^2 + y^2, \quad f_2(x, y) = -x^2 - y^2, \quad f_3(x, y) = x^2 - y^2.$$

As polynomials, these three functions are twice continuously differentiable. One further computes that

$$f_i(0, 0) = 0, \quad \nabla f_i(0, 0) = 0, \quad \text{for } i \in \{1, 2, 3\}$$

[6] This awkward word is derived from the name of the German mathematician Otto Hesse (1811-1874).

and that

$$H_{f_1}(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad H_{f_2}(0,0) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}, \qquad H_{f_3}(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}.$$

As all three matrices are diagonal matrices, their eigenvalues are precisely the diagonal entries. Using Theorem 12.64 we derive that $H_{f_1}(0,0)$ is positive definite, $H_{f_2}(0,0)$ is negative definite and $H_{f_3}(0,0)$ is indefinite. Hence, f_1 has a local minimum in $(0,0)$, f_2 has a local maximum in $(0,0)$ and f_3 has neither of the two. (See the lecture videos for an illustration.)

(2) If $H_f(a)$ satisfies none of the three definiteness properties, then a might be a local extremum point or it might not, but it can not be determined from $H_f(a)$ alone. As an example, consider the three functions

$$f_1, f_2, f_3 : \mathbb{R}^2 \to \mathbb{R}, \qquad f_1(x,y) = x^2 + y^4, \qquad f_2(x,y) = x^2 - y^4, \qquad f_3(x,y) = x^2 + y^3$$

Then $f_i(0,0) = 0$ and $\nabla f_i(0,0) = (0,0)$ for all $i \in \{1,2,3\}$ and one further checks that

$$H_{f_i}(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \quad \forall i \in \{1,2,3\}.$$

Evidently, this matrix has 2 and 0 as eigenvalues, so it is neither positive definite nor negative definite nor indefinite. But even as gradient and Hessian coincide, the three functions show entirely different behaviors near $(0,0)$:

(i) Since evidently $f_1(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$, f_1 attains its minimum 0 in $(0,0)$.

(ii) Since $f_2(t,0) \geq 0$ and $f_2(0,t) \leq 0$ for all $t \in \mathbb{R}$, it follows that f_2 does not have a local extremum in $(0,0)$. However, similar to the situation in part (1) of this remark, $f_2(\cdot,0) : \mathbb{R} \to \mathbb{R}$, $f_2(x,0) = x^2$, attains its mininum in 0, while $f_2(0,\cdot) : \mathbb{R} \to \mathbb{R}$, $f_2(0,y) = -y^4$ attains its maximum in 0.

(iii) Since $f_3(t,0) \geq 0$ for all $t \in \mathbb{R}$, but $f_3(0,t) \leq 0$ for all $t \in (-\infty, 0)$, f_3 does not have a local extremum in $(0,0)$. But in contrast to the situation for f_2, the map $f_3(0,\cdot) : \mathbb{R} \to \mathbb{R}$, $f_3(0,y) = y^3$, does not have a local extremum in 0 either.

As the reader might guess, similar to the single-variable setting, one might in such cases employ the differentials $D^k f_a$ for $k \geq 3$ to make further statements about the local extrema, but as this approach becomes rather complicated, we refrain from going into detail.

Example 12.72. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = (x^2 + x - 2)^2 + xy^2$. f is a polynomial, thus it is twice differentiable. We compute for all $(x,y) \in \mathbb{R}^2$ that

$$\partial_1 f(x,y) = 2(x^2 + x - 2)(2x + 1) + y^2, \qquad \partial_2 f(x,y) = 2xy.$$

Hence,

$$\nabla f(x,y) = 0 \quad \Leftrightarrow \quad \begin{cases} 2(x^2 + x - 2)(2x + 1) + y^2 & = 0, \\ 2xy & = 0. \end{cases}$$

By the second equation, $x = 0$ or $y = 0$. If $x = 0$, then the first equation becomes $-4 + y^2 = 0$, hence $y = 2$ or $y = -2$. If $y = 0$, then the first equation yields

$$(x^2 + x - 2)(2x + 1) = 0 \quad \Leftrightarrow \quad x^2 + x - 2 = 0 \ \vee x = -\tfrac{1}{2}.$$

The first equation has the two solutions $x = 1$ and $x = -2$. Thus, we have altogether shown that

$$\operatorname{Crit} f = \{(0,2), (0,-2), (-\tfrac{1}{2}, 0), (1,0), (-2,0)\}.$$

To check which of the critical points are local maximum points or local minimum points, we need to determine the Hessian of f in these points. We compute for all $(x,y) \in \mathbb{R}^2$ that

$$H_f(x,y) = \begin{pmatrix} \partial_1\partial_1 f(x,y) & \partial_1\partial_2 f(x,y) \\ \partial_2\partial_1 f(x,y) & \partial_2\partial_2 f(x,y) \end{pmatrix} = \begin{pmatrix} 12x^2 + 12x - 6 & 2y \\ 2y & 2x \end{pmatrix}.$$

For the critical points of f, we obtain

$$H_f(0,2) = \begin{pmatrix} -6 & 4 \\ 4 & 0 \end{pmatrix}, \quad H_f(0,-2) = \begin{pmatrix} -6 & -4 \\ -4 & 0 \end{pmatrix}, \quad H_f(-\tfrac{1}{2},0) = \begin{pmatrix} -9 & 0 \\ 0 & -1 \end{pmatrix},$$
$$H_f(1,0) = \begin{pmatrix} 18 & 0 \\ 0 & 2 \end{pmatrix}, \quad H_f(-2,0) = \begin{pmatrix} 18 & 0 \\ 0 & -4 \end{pmatrix}.$$

Using the criteria from Corollary 12.65, we conclude that f has a local maximum in $(-\frac{1}{2}, 0)$ and a local minimum in $(1,0)$ and no local extrema in the other three critical points. The values of the local maximum and minimum are given by $f(-\frac{1}{2}, 0) = \frac{25}{16}, \qquad f(1,0) = 0.$

With the results of this section, we are now able to give a general method for finding maxima and minima of differentiable functions on compact sets. Let $K \subset \mathbb{R}^n$ be compact. Then by Corollary 11.77, every continuous function $f : K \to \mathbb{R}$ attains its maximum and its minimum, but that theorem and its prove provide no constructive method of finding their values. Consider the interior $\mathring{K}$ of K. By Theorem 11.59.a), $\mathring{K}$ is open and if f attains its maximum (or minimum) in a point in $\mathring{K}$, then it particularly has a local maximum (or minimum) in that point. So if $f|_{\mathring{K}} : \mathring{K} :\to \mathbb{R}$ is twice differentiable, we can apply Theorem 12.68 to find such points. In general, one can in many cases use the following method to determine maxima and minima.

Method for finding local and global extrema of twice continuously differentiable functions on compact sets:

Let $n \in \mathbb{N}$ and let $K \subset \mathbb{R}^n$ be compact. Let $f : K \to \mathbb{R}$ be continuous and assume that $f|_{\mathring{K}} : \mathring{K} \to \mathbb{R}$ is twice continuously differentiable.

- *Necessary condition for local extrema:* Find all $a \in \mathring{K}$ with $\nabla f(a) = 0$, i.e. $\partial_i f(a) = 0$ for all $i \in \{1, 2, \dots, n\}$. (If there are no such points, f has no local extrema in $\mathring{K}$ and attains its maximum in ∂K. Then one can skip the next step.)
- *Sufficient conditions for local extrema:* For each $a \in U$ with $\nabla f(a) = 0$ compute $H_f(a)$.
  - If $H_f(a)$ does not have 0 as an eigenvalue, use Theorem 12.68 to decide whether a is a local maximum point, a local maximum point or neither of the two.
  - If $H_f(a)$ has 0 as an eigenvalue, then try to decide if it is a local extremum point by other methods (e.g. show explicitly that for each $r > 0$ there are $x, y \in B_r(a)$ with $f(x) < f(a) < f(y)$ to derive that f does not have a local extremum in a.)
- To find the maximum of f, proceed as follows (the minimum is found analogously):
  - Compute $M_1 := \max\{f(a) \mid a \in \mathring{K},\ f \text{ has a local maximum in } a\}$ from the results of the previous steps (analogously for the minimum).
  - Compute $M_2 := \max\limits_{x \in \partial K} f(x)$ by explicit computations .

  Then $M := \max\{M_1, M_2\}$ is the maximum of f.

Example 12.73. Consider $A := \{(x, y) \in \mathbb{R}^2 \mid |x| + |y| \leq 1\}$ and $f : A \to \mathbb{R}$, $f(x, y) = x^2 + xy + y^2$. Since every $(x, y) \in A$ satisfies $\|(x, y)\| = \sqrt{x^2 + y^2} \leq \sqrt{x^2} + \sqrt{y^2} = |x| + |y| \leq 1$, it follows that $A \subset \overline{B}_1(0, 0)$, hence A is bounded.
It further holds that $A = g^{-1}((-\infty, 1])$, where $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x, y) = |x| + |y|$. The map g is continuous and we compute that

$$\mathbb{R}^2 \setminus A = g^{-1}((1, +\infty)),$$

which is open by Theorem 11.51. Hence, A is closed. By the Heine-Borel theorem, A is just compact. Since f is continuous as a polynomial, it attains its maximum and minimum on A by Corollary 11.77 and we want to compute their values.
We first determine the local extrema of f in $\mathring{A} = \{(x, y) \in \mathbb{R}^2 \mid |x| + |y| < 1\}$. We compute that

$$\partial_1 f(x, y) = 2x + y, \qquad \partial_2 f(x, y) = x + 2y \qquad \forall (x, y) \in \mathbb{R}^2.$$

Thus any local maximum or minimum point satisfies $\begin{cases} 2x + y &= 0, \\ x + 2y &= 0, \end{cases}$ and one easily checks that the unique solution of this system is $(0, 0)$. The Hessian of f in an arbitrary point computes as $H_f(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Since $\det \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 4 - 3 = 1 > 0$ and $2 + 2 = 4 > 0$, it follows from Corollary 12.65.a) that $H_f(0, 0)$ is positive definite, hence f has a local minimum in $(0, 0)$, which is given by $f(0, 0) = 0$.

For $(x,y) \in A$ with $xy \geq 0$, one easily sees that $f(x,y) \geq 0$. For $(x,y) \in A$ with $xy \leq 0$ we compute that

$$f(x,y) = x^2 + xy + y^2 \geq x^2 + 2xy + y^2 = (x+y)^2 \geq 0.$$

Hence, 0 is the minimum of f, which is attained in $(0,0)$.
Since f has no local maximum points in $\mathring{A}$, it further attains its maximum in $\partial A = \{(x,y) \in \mathbb{R}^2 \mid |x| + |y| = 1\}$. For $(x,y) \in \partial A$ we compute that

$$f(x,y) = x^2 + xy + y^2 \leq x^2 + |xy| + y^2 \leq |x|^2 + 2|xy| + |y|^2 = (|x| + |y|)^2 = 1.$$

Since $f(1,0) = 1$, it follows that the maximum of f is 1 and one checks that it is attained in $(1,0)$, $(-1,0)$, $(0,1)$ and $(0,-1)$.

Chapter 13

The implicit function theorem and its friends

This chapter will not be part of the exam or the retake exam.

This chapter has a slightly different structure than the previous ones. As a climax of multivariable calculus, we will discuss several "big" and important theorems from calculus that are invaluable for applications in all kinds of calculus and geometry.
Most applications of mathematics in the sciences eventually leave the applicant with the same task, in one way or the other: *find the solution sets of certain equations*. In the linear algebra part of this course, we have learnt how to find out if systems of linear equations have solutions and to compute their solution sets. However, most of the equations that appear in applications of mathematics, will not just be given as linear equations. For example, many physical laws like conservation of energy are given as equations of the form

$$f(x) = y,$$

where f is some (possibly very complicated) continuous or differentiable function. An example is given by the following system of equations which we will encounter again in Section 13.2.

$$\begin{cases} y^2 + x^2 \cos y - \sin(x^2) & = u, \\ \sin x + xy & = v, \end{cases}$$

Given $u, v \in \mathbb{R}$ as fixed values, is there a solution of this system of equations? So far, we have no methods to tackle such a problem.

In each of the following sections, we will motivate one of the big theorems by taking a look at a problem involving solutions of non-linear equations. We will then show in the end of each section how each of the problems can be resolved by applying the theorem discussed in that section.

13.1 The Banach fixed point theorem

Let us consider the following problem:

Is there a bounded function $f : \mathbb{R} \to \mathbb{R}$ *which satisfies*

$$f(2x) = 2f(x) - \sin x \quad \forall x \in \mathbb{R} \quad ?$$

If so, how many functions satisfy this equation?

So far, there is no obvious way to find the answer to that problem. This equation is not an ordinary differential equation (for which there are many solution methods) and finding such a solution for a fixed x will not help as such a function is unlikely to solve the equation for all $x \in \mathbb{R}$.

We will trace this question back to the next theorem, which has loads of applications throughout mathematics. For us it will be an important tool to prove the inverse function theorem and the implicit function theorem.

Theorem 13.1 (Banach fixed point theorem[1])**.** *Let (X, d) be a complete metric space. Let $f : X \to X$ be given, such that there exists some $C \in [0, 1)$, such that*

$$d(f(x), f(y)) \leq C \cdot d(x, y) \qquad \forall x, y \in X.$$

(A map with this property is also called a contraction of X*.) Then f has a unique fixed point, i.e. there exists a unique $x_* \in X$ with $f(x_*) = x_*$.*

Proof. (1) *Uniqueness:* Assume there are $x_1, x_2 \in X$ with $f(x_1) = x_1$, $f(x_2) = x_2$. Then

$$d(x_1, x_2) = d(f(x_1), f(x_2)) \leq C \cdot d(x_1, x_2)$$

and thus $(1 - C)d(x_1, x_2) \leq 0$. Since $1 - C > 0$ and $d(x_1, x_2) \geq 0$, this yields that $d(x_1, x_2) = 0$ and thus $x_1 = x_2$ by property (i) of a metric.

(2) *Existence:* Let $x_0 \in X$ be arbitrary and consider the sequence $(x_n)_{n \in \mathbb{N}}$, which is recursively defined by

$$x_1 := x_0, \quad x_{n+1} := F(x_n) \quad \forall n \in \mathbb{N}.$$

Then for all $n \geq 2$ it holds that

$$d(x_{n+1}, x_n) = d(F(x_n), F(x_{n-1})) \leq C \cdot d(x_n, x_{n-1}).$$

Iterating this inequality, one shows for all $n \in \mathbb{N}$ that

$$d(x_{n+1}, x_n) \leq C \cdot d(x_n, x_{n-1}) \leq C^2 d(x_{n-1}, x_{n-2}) \leq \cdots \leq C^{n-1} d(x_2, x_1). \tag{13.1}$$

We put $a := d(x_2, x_1)$. For all $m, n \in \mathbb{N}$ with $m < n$ we compute using the triangle inequality that

$$\begin{aligned} d(x_n, x_m) &\leq d(x_n, x_{n-1}) + d(x_{n-1}, x_m) \leq \cdots \leq \sum_{k=1}^{n-m} d(x_{m+k}, x_{m+k-1}) \\ &\overset{(13.1)}{\leq} \sum_{k=1}^{n-m} C^{m+k-2} a = C^{m-1} \sum_{k=0}^{n-m-1} C^k a \\ &\leq C^{m-1} \sum_{k=0}^{\infty} C^k a = \frac{C^{m-1}}{1 - C}, \end{aligned}$$

[1] This theorem is also called the *contraction principle* or *contraction mapping theorem* in various textbooks.

where we used the convergence and the sum formula of the geometric series and the fact that $C < 1$.

Given $\varepsilon > 0$, there exists an $n_0 \in \mathbb{N}$, such that $C^{n_0-1} < (1-C)\varepsilon$. By the previous computation, this yields that $d(x_n, x_m) < \varepsilon$ for all $m, n \in \mathbb{N}_0$. Hence, $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since X is complete, $(x_n)_{n\in\mathbb{N}}$ is convergent and we put $x_* := \lim_{n\to\infty} x_n$. Since f is by assumption Lipschitz-continuous, it is continuous by Theorem 11.39. Since any subsequence of $(x_n)_{n\in\mathbb{N}}$ converges against the same limit, we compute from the definition of $(x_n)_{n\in\mathbb{N}}$ that

$$x_* = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n) \overset{\text{Theorem 11.30}}{=} f\left(\lim_{n\to\infty} x_n\right) = f(x_*).$$

Hence, x_* is a fixed point of f, as we wanted to show.

□

Remark 13.2. The Banach fixed point theorem also gives us a method at hand of how to find the fixed point of a contraction $f : X \to X$, where X is a complete metric space: if $x_0 \in X$ is chosen arbitrarily, then the fixed point is given by

$$x_* = \lim_{n\to\infty} f^n(x_0),$$

where $f^n = f \circ f \circ \cdots \circ f$ denotes the n-fold composition of f with itself.

Example 13.3. We reconsider the equation from the beginning of the section,

$$f(2x) = 2f(x) - \sin x \quad \forall x \in \mathbb{R}, \tag{13.2}$$

where $f : \mathbb{R} \to \mathbb{R}$ is bounded. Equivalently, $f(x) = \frac{1}{2}(f(2x) + \sin x)$ for all $x \in \mathbb{R}$ which gives us the idea for the connection to the Banach fixed point theorem.
By Theorem 11.17, $(\mathcal{B}(\mathbb{R}, \mathbb{R}), \|\cdot\|_\infty)$ is a Banach space, hence a complete metric space. Consider the map

$$F : \mathcal{B}(\mathbb{R}, \mathbb{R}) \to \mathcal{B}(\mathbb{R}, \mathbb{R}), \quad (F(f))(x) = \frac{1}{2}(f(2x) + \sin x).$$

Then $f \in \mathcal{B}(\mathbb{R}, \mathbb{R})$ satisfes (13.2) if and only if f is a fixed point of F. For all $f, g \in \mathcal{B}(\mathbb{R}, \mathbb{R})$ we compute that

$$\begin{aligned}
\|F(f) - F(g)\|_\infty &= \sup_{x\in\mathbb{R}} |(F(f))(x) - (F(g))(x)| \\
&= \sup_{x\in\mathbb{R}} \left|\frac{1}{2}f(2x) - \frac{1}{2}g(2x)\right| = \frac{1}{2}\sup_{x\in\mathbb{R}} |f(2x) - g(2x)| = \frac{1}{2}\sup_{x\in\mathbb{R}} |f(x) - g(x)| \\
&= \frac{1}{2}\|f - g\|_\infty.
\end{aligned}$$

In particular, $\|F(f) - F(g)\|_\infty \leq \frac{1}{2}\|f - g\|_\infty$ for all $f, g \in \mathcal{B}(\mathbb{R}, \mathbb{R})$, so by the Banach fixed point theorem F has a unique fixed point and thus (13.2) *has a unique solution.*
One can further show that the unique solution f_* is *continuous*: if we define a sequence $(f_n)_{n\in\mathbb{N}}$ by apply the method outlined in Remark 13.2 to the zero function, i.e.

$$f_1(x) = (F(0))(x) = \frac{1}{2}\sin x, \quad f_2(x) = (F(f_1))(x) = \frac{1}{4}\sin(2x) + \frac{1}{2}\sin x, \quad \ldots,$$

then f_n is continuous for each $n \in \mathbb{N}$. Since $f_* = \lim_{n\to\infty} f_n$ is the uniform limit of $(f_n)_{n\in\mathbb{N}}$, it follows from Theorem 11.23 that f_* is again continuous.

13.2 The inverse function theorem

Consider the following system of equations

$$\begin{cases} y + x^2 \cos y - \sin(x^2) & = u, \\ \sin x + xy & = v, \end{cases} \tag{13.3}$$

where we view $u, v \in \mathbb{R}$ as given and $x, y \in \mathbb{R}$ are variables. Given $u, v \in \mathbb{R}$, are there $x, y \in \mathbb{R}$ which solve this system? We have no explicit solution methods and no general approach to such a system.
Taking a closer look and trying out some numbers, one sees that for the case of $u = v = 0$, we obtain $x = 0$ and $y = 0$ as a solution of the system. Since all of the functions involved are differentiable, hence continuous, it seems reasonable to guess that for u and v close to 0, should have a solution as well. More precisely, one might ask the following question:

If $V \subset \mathbb{R}^2$ is a sufficently small neighborhood of $(0,0)$, does the system (13.3) *have a solution for each $(u,v) \in V$?*

We will answer this question by applying an important result from calculus that we are going to prove in this section.
Let $U, V \subset \mathbb{R}^n$ be open and assume that $f : U \to V$ is a bijection with inverse $f^{-1} : V \to U$. If f is differentiable in $a \in U$ and f^{-1} is differentiable in $b := f(a) \in V$, what can we say about their differentials?
The chain rule applied to $f \circ f^{-1} = \mathrm{id}_V$ and $f^{-1} \circ f = \mathrm{id}_U$ yields

$$Df_a \circ (Df^{-1})_b = \mathrm{id}_{\mathbb{R}^n} \quad \text{and} \quad (Df^{-1})_b \circ Df_a = \mathrm{id}_{\mathbb{R}^n}.$$

This shows that Df_a is invertible and that its inverse is given by $(Df^{-1})_b$. Hence,

$$(Df^{-1})_b = (Df_a)^{-1}.$$

However, this equation by itself is not that useful, since the assumptions are quite strong. Given $f : U \to V$, we already need to know that f^{-1} is differentiable at the respective point. (We have already seen in Mathematics 1 that this is not always the case, consider for example the case of $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$.) Since f^{-1} might be hard to determine explicitly, this is a big problem for applications to concrete maps. The following result gives a criterion which ensures the differentiability of the inverse map.

Theorem 13.4. *Let $n \in \mathbb{N}$, $U, V \subset \mathbb{R}^n$ be open and let $f : U \to V$ be bijective. Assume that f is differentiable in $a \in U$ and that $Df_a \in L(\mathbb{R}^n, \mathbb{R}^n)$ is an isomorphism. If $f^{-1} : V \to U$ is continuous in $b := f(a)$, then f^{-1} is differentiable in b with*

$$(Df^{-1})_b = (Df_a)^{-1}.$$

Proof. Let $U' := \{h \in \mathbb{R}^n \mid a + h \in U\}$ and $V' := \{h \in \mathbb{R}^n \mid b + h \in V\}$. To show the claim, we need to prove that

$$\lim_{h \to 0} \frac{1}{\|h\|} \left(f^{-1}(b+h) - f^{-1}(b) - (Df_a)^{-1}(h) \right) \stackrel{!}{=} 0,$$

or equivalently,

$$\lim_{h \to 0} \frac{\|f^{-1}(b+h) - f^{-1}(b) - (Df_a)^{-1}(h)\|}{\|h\|} \stackrel{!}{=} 0. \tag{13.4}$$

Put $r(h) := f^{-1}(b+h) - f^{-1}(b) = f^{-1}(b+h) - a$. Then $a + r(h) = f^{-1}(b+h)$, so applying f to both sides of the equation yields $b + h = f(a + r(h))$. Since f is differentiable in a, there exists $\varphi : U' \to \mathbb{R}^n$ with $\lim_{h\to 0} \frac{1}{\|h\|}\varphi(h) = 0$, such that for all $h \in V'$:

$$\begin{aligned} f(a+r(h)) = f(a) + Df_a(r(h)) + \varphi(r(h)) \quad &\Leftrightarrow \quad b + h = b + Df_a(r(h)) + \varphi(r(h)) \\ &\Leftrightarrow \quad h = Df_a(r(h)) + \varphi(r(h)). \end{aligned}$$

Applying the linear map $(Df_a)^{-1}$ to both sides of this last equation yields

$$(Df_a)^{-1}(h) = r(h) + (Df_a)^{-1}(\varphi(r(h))). \tag{13.5}$$

We further compute that

$$f^{-1}(b+h) - f^{-1}(b) - (Df_a)^{-1}(h) = r(h) - (Df_a)^{-1}(h) = Df_a^{-1}(\varphi(r(h)))$$

and thus

$$\lim_{h\to 0} \frac{\|f^{-1}(b+h) - f^{-1}(b) - (Df_a)^{-1}(h)\|}{\|h\|} = \lim_{h\to 0} \frac{\|(Df_a)^{-1}(\varphi(r(h))\|}{\|h\|}.$$

By Corollary 11.42 and Theorem 11.44, there exists $C > 0$, such that

$$\|Df_a^{-1}(v)\| \leq C\|v\| \qquad \forall v \in \mathbb{R}^n.$$

Inserting this into the above limit yields

$$\lim_{h\to 0} \frac{\|f^{-1}(b+h) - f^{-1}(b) - (Df_a)^{-1}(h)\|}{\|h\|} \leq C \lim_{h\to 0} \frac{\|\varphi(r(h))\|}{\|h\|} = \lim_{h\to 0} \frac{\|\varphi(r(h))\|}{\|r(h)\|} \cdot \frac{\|r(h)\|}{\|h\|}.$$

Since $\lim_{h\to 0} \frac{\|\varphi(r(h))\|}{\|r(h)\|} = 0$ by the defining property of φ, it suffices to show that the quotient $\frac{\|r(h)\|}{\|h\|}$ is bounded in a neighborhood of 0 to obtain that the rightmost expression tends to 0.. From (13.5) we obtain that

$$\|r(h)\| = \|Df_a^{-1}(h - \varphi(r(h)))\| \leq C\|h - \varphi(r(h))\| \leq C(\|h\| + \|\varphi(r(h))\|).$$

Since $\lim_{h\to 0} \frac{\|\varphi(h)\|}{\|h\|} = 0$, there exists $\varepsilon > 0$ with $B_\varepsilon(0) \subset U'$, such that $\|\varphi(h)\| \leq \frac{1}{2C}\|h\|$ for all $h \in B_\varepsilon(a)$. Since f^{-1} is continuous in b, it holds that $\lim_{h\to 0} r(h) = 0$. Thus, there exists $\delta > 0$ with $B_\delta(0) \subset V'$, such that $r(B_\delta(0)) \subset B_\varepsilon(0)$. Hence, for all $h \in B_\delta(0)$ we obtain

$$\|r(h)\| \leq C(\|h\| + \|\varphi(r(h))\|) \leq C\|h\| + \frac{1}{2}\|r(h)\| \quad \Leftrightarrow \quad \|r(h)\| \leq 2C\|h\|.$$

Hence, $\frac{\|r(h)\|}{\|h\|} \leq 2C$ for all $h \in B_\delta(0)$. Hence, $\lim_{h\to 0} \frac{\|\varphi(r(h))\|}{\|r(h)\|} \cdot \frac{\|r(h)\|}{\|h\|} = 0$, which implies (13.4) and thereby completes the proof. □

The previous theorem has some flaws that makes it inconvenient for applications. Given a differentiable map $f : U \to V$ between open subsets of $\mathbb{R}^n$, we can compute its Jacobi matrix $J_f(a)$ at some $a \in U$, check whether it is invertible by our knowledge from linear algebra and conclude if the differentiable is an isomorphism or not. However, given a complicated differentiable map it is not at all obvious to see if the map is bijective. One needs to find out if the

equation $f(x) = y$ has a unique solution $x \in U$ for every $y \in V$, for which there is no general method.
This problem is tackled by the inverse function theorem. By using a lot of our knowledge on multivariable calculus and by using the Banach fixed point theorem in a very non-obvious way, we will prove that thoerem, which shows us that the bijectivity of a map between some "small" open sets can already be derived if one observes that the differential of the map is an isomorphism.

Theorem 13.5 (Inverse Function Theorem)**.** *Let $n \in \mathbb{N}$, $W \subset \mathbb{R}^n$ be open, $f : W \to \mathbb{R}^n$ be continuously differentiable and $a \in W$. If $Df_a : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism, then there is an open neighborhood $U \subset W$ of a, such that*

(i) $f : U \to V$, where $V := f(U)$, is bijective and V is open,

(ii) $Df_x : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism for each $x \in U$,

(iii) $f^{-1} : V \to U$ is continuously differentiable with $(Df^{-1})_{f(x)} = (Df_x)^{-1}$ for all $x \in U$.

Proof. We begin this long proof by making some simplifying assumptions:

- We assume w.l.o.g. that $Df_a = \mathrm{id}_{\mathbb{R}^n}$: if this is not the case, then we can replace f by the map $F : W \to \mathbb{R}^n$, $F := (Df_a)^{-1} \circ f$. One derives from Theorem 12.22 and the chain rule that in this case
$$DF_a = Df_a^{-1} \circ Df_a = \mathrm{id}_{\mathbb{R}^n}.$$
Since Df_a^{-1} is an isomorphism, one checks that F has properties (i) and (ii) if and only if f has properties (i) and (ii).
- We assume w.l.o.g. that $a = 0$ and $f(a) = 0$: if this is not the case, then we can replace f by the map $G : W' \to \mathbb{R}^n$, $G(x) = f(a + x) - f(a)$, where $W' := \{x \in \mathbb{R}^n \mid a + x \in W\}$. One checks that f has properties (i) and (ii) with respect to a if and only if G has properties (i) and (ii) with respect to 0.

So in the following we assume that $a = 0$, $f(0) = 0$ and $Df_a = \mathrm{id}_{\mathbb{R}^n}$.

(1) We first want to show that there is a compact neighborhood K of 0 in W, such that Df_x is an isomorphism for all $x \in K$.

Since f is continuously differentiable and $Df_0 = \mathrm{id}_{\mathbb{R}^n}$, it holds that
$$\lim_{x \to a} \partial_i f(x) = \partial_i f(0) \qquad \forall i \in \{1, 2, \dots, n\}.$$
Thus, there exists $r > 0$ with $K := \overline{B}_{2r}(0) \subset W$, such that
$$\|\partial_i f(0) - \partial_i f(x)\| \leq \frac{1}{2n^{\frac{3}{2}}} \qquad \forall x \in \overline{B}_{2r}(0),\ i \in \{1, 2, \dots, n\}.$$
For all $x \in K$ and $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ we thus compute that
$$\begin{aligned} \|v - Df_x(v)\| = \|Df_0(v) - Df_x(v)\| &= \Big\| \sum_{i=1}^n \partial_i f(0) v_i - \sum_{i=1}^n \partial_i f(x) v_i \Big\| \\ &\leq \sum_{i=1}^n \|\partial_i f(0) - \partial_i f(x)\| \, |v_i| \leq \frac{1}{2n^{\frac{3}{2}}} \sum_{i=1}^n |v_i| \leq \frac{1}{2n^{\frac{3}{2}}} n \|v\| \\ \Rightarrow \quad \|v - Df_x(v)\| &\leq \frac{1}{2\sqrt{n}} \|v\|. \end{aligned} \tag{13.6}$$

We want to derive that Df_x is an isomorphism for all $x \in K$. By Corollary 7.23, it suffices to show that $\ker Df_x = \{0\}$ for each $x \in K$. If $v \in \ker Df_x$, then

$$\|v\| = \|v - Df_x(v)\| \overset{(13.6)}{\leq} \frac{1}{2\sqrt{n}}\|v\|.$$

But since $\frac{1}{2\sqrt{n}} < 1$, this can only be satisfied for $v = 0$. Hence Df_x is an isomorphism for all $x \in K$.

(2) Next we want to show that there is a neighborhood $V \subset \mathbb{R}^n$ of 0, such that for all $y \in V$ the equation $f(x) = y$ has a solution $x \in K$ with K given as in the previous step.

Our strategy is to turn this into a "fixed point problem" which we will solve using the Banach fixed point problem. For a given $y \in \mathbb{R}^n$ let

$$\phi_y : W \to \mathbb{R}^n, \qquad \phi_y(x) := y + x - f(x).$$

Then $f(x) = y$ holds if and only if $\phi_y(x) = x$. Since f is continuously differentiable, ϕ_y is continuously differentiable and from Theorem 12.23.a) we obtain

$$(D\phi_y)_x = \mathrm{id}_{\mathbb{R}^n} - Df_x \qquad \forall x \in W.$$

For each $x \in K$ we thus compute from (13.6) that

$$\|(D\phi_y)_x(v)\| \leq \frac{1}{2\sqrt{n}}\|v\| \quad \forall v \in \mathbb{R}^n.$$

Let $\phi_y = (\phi_1, \ldots, \phi_n)$, where $\varphi_i : W \to \mathbb{R}$ for each $i \in \{1, 2, \ldots, n\}$. Since K is convex, we obtain from the mean value theorem (Theorem 12.37) that for all $x_1, x_2 \in K$ and all $i \in \{1, 2, \ldots, n\}$ there exists $\xi_i \in K$, such that

$$|\phi_i(x_1) - \phi_i(x_2)| = |(D\phi_i)_{\xi_i}(x_1 - x_2)| \leq \|(D\phi_y)_{\xi_i}(x_1 - x_2)\| \leq \frac{1}{2\sqrt{n}}\|x_1 - x_2\|.$$

Consequently,

$$\big\|\phi_y(x_1) - \phi_y(x_2)\big\|^2 = \sum_{i=1}^{n} |\phi_i(x_1) - \phi_i(x_2)|^2 \leq \sum_{i=1}^{n} \frac{1}{4n}\|x_1 - x_2\|^2 = \frac{1}{4}\|x_1 - x_2\|^2$$

and thus

$$\|\phi_y(x_1) - \phi_y(x_2)\| \leq \frac{1}{2}\|x_1 - x_2\| \qquad \forall x_1, x_2 \in K. \tag{13.7}$$

For all $x \in K$ and all $y \in B_r(0)$ we compute that

$$\|\phi_y(x)\| \leq \|\phi_y(x) - \phi_y(0)\| + \|\phi_y(0)\| \leq \frac{1}{2}\|x - 0\| + \|y\| = \frac{1}{2}\|x\| + \|y\| < 2r.$$

and thus $\phi_y(K) \subset B_{2r}(0)$. Hence, for $y \in B_r(0)$, the map ϕ_y restricts to a map $\phi_y|_K : K \to K$. Since K is closed, one checks (this is left to the reader as an exercise) that K with the restriction of the Euclidean metric is a complete metric space. By (13.7), each $\phi_y|_K$ is a contraction, so by the Banach fixed point theorem, for each $y \in B_r(0)$ there exists a unique $x \in K$ with $\phi_y(x) = x$ and thus $f(x) = y$. Since $\phi_y(K) \subset B_{2r}(0)$, it further holds that $x \in B_{2r}(0)$.

(3) Put $V := B_r(0)$, which is an open neighborhood of 0. Since f is continuous and $f(0) = 0$, $f^{-1}(V)$ is an open neighborhood of 0 by Theorem 11.51, hence $U := B_{2r}(0) \cap f^{-1}(V)$ is an open neighborhood of 0 by Theorem 11.49.c). By the results of the previous step, $f(U) \subset V$ and for each $y \in V$ there is a unique $x_y \in U$ with $f(x_y) = y$. Hence, the restriction $f|_U : U \to V$ is a bijection.

(4) We want to show that the inverse of $f|_U$, which we simply denote by $f^{-1} : V \to U$, is continuous. Let $y_1, y_2 \in V$ and put $x_1 := f^{-1}(y_1)$ and $x_2 := f^{-1}(y_2)$. Then with $y = 0$, it follows from (13.7) using that $\phi_0(x) = x + f(x)$ that

$$\begin{aligned} \|f^{-1}(y_1) - f^{-1}(y_2)\| = \|x_1 - x_2\| &= \|\phi_0(x_1) - \phi_0(x_2) - (f(x_1) - f(x_2)\| \\ &\leq \|\phi_0(x_1) - \phi_0(x_2)\| + \|f(x_1) - f(x_2)\| \\ &\leq \frac{1}{2}\|x_1 - x_2\| + \|y_1 - y_2\| \\ \Rightarrow \quad \|f^{-1}(y_1) - f^{-1}(y_2)\| &\leq \frac{1}{2}\|f^{-1}(y_1) - f^{-1}(y_2)\| + \|y_1 - y_2\|. \end{aligned}$$

This yields that $\|f^{-1}(y_1) - f^{-1}(y_2)\| \leq 2\|y_1 - y_2\|$ for all $y_1, y_2 \in V$. Thus, f^{-1} is Lipschitz-continuous, hence continuous by Theorem 11.39.

(5) Since $f : U \to V$ is bijective and continuously differentiable with Df_x an isomorphism for each $x \in U$ and since $f^{-1} : V \to U$ is continuous, it follows from Theorem 13.4 that f^{-1} is differentiable with $(Df^{-1})_{f(x)} = (Df_x)^{-1}$ for all $x \in U$.

□

Remark 13.6. In the situation of the inverse function theorem, we obtain from that theorem and from the definition of inverse matrices that the Jacobi matrices of f and f^{-1} satisfy

$$J_{f^{-1}}(b) = (J_f(a))^{-1}.$$

This particularly means that *without explicitly knowing* f^{-1} we can compute the differential of f^{-1} by inverting the Jacobi matrix of f in the respective points. Here, we can use the matrix inversion methods that we learnt about in Chapter 8, so we can find the differential of f^{-1} by purely algebraic methods. Isn't that lovely?

Example 13.7. We reconsider the system of equations (13.3). If we put

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \qquad f(x, y) = (y + x^2 \cos y - \sin(x^2), \sin x + xy),$$

then $(x, y) \in \mathbb{R}^2$ is a solution of (13.3) if and only if $f(x, y) = (u, v)$.
We have observed above that $f(0, 0) = (0, 0)$. f is obviously continuously differentiable with

$$J_f(x, y) = \begin{pmatrix} 2x(\cos y - \cos x^2) & 1 - x^2 \sin y \\ \cos x + y & x \end{pmatrix} \quad \forall (x, y) \in \mathbb{R}^2.$$

In particular, $J_f(0, 0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, so that $\det J_f(0, 0) = -1 \neq 0$, hence $J_f(0, 0)$ is invertible. This yields that $Df_{(0,0)}$ is an isomorphism.
Thus, by the inverse function theorem, there are neighborhoods U and V of $(0, 0)$ such that for each $(u, v) \in V$ there is a unique $(x, y) \in U$ with $f(x, y) = (u, v)$. In other words, for each $(u, v) \in V$, the system (13.3) has a unique solution in U.

We can go even one step further and try to approximate the solution of (13.3) for $(u,v) \in V$ using the methods we developed, namely by Taylor's formula. Fix $(u,v) \in V$. Since $f^{-1} : V \to U$ is differentiable, we obtain by Taylor's theorem applied to $a = (0,0)$ that there is a map $\varphi : V \to \mathbb{R}^2$ with $\lim_{(u,v)\to 0} \frac{\|\varphi(u,v)\|}{\|(u,v)\|} = 0$, such that

$$\begin{aligned} f^{-1}(u,v) &= f^{-1}(0,0) + (Df^{-1})_{(0,0)}(u,v) + \varphi(u,v) \\ &= 0 + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \varphi(u,v) \\ &= (v,u) + \varphi(u,v). \end{aligned}$$

Hence, for each $(u,v) \in V$ the unique solution of (13.3) in U satisfies

$$\begin{aligned} x(u,v) &= v + \varphi_1(u,v), \\ y(u,v) &= u + \varphi_2(u,v), \end{aligned}$$

where $\varphi = (\varphi_1, \varphi_2)$. From the limiting properties of φ one could then compute approximate solutions of the system.

13.3 The implicit function theorem

Let us consider the equation

$$x^2 + y^2 = 1,$$

where $x, y \in \mathbb{R}$. The solution set of this equation as a subset of $\mathbb{R}^2$ is precisely the unit circle, the circle of radius 1 centered in the origin. While this circle which we will denote by K is a subset of $\mathbb{R}^2$ it "looks one-dimensional" in nature in that a small piece of K looks like a bent or curved line segment. We want to study this observation in greater detail.
Let $(x_0, y_0) \in K$ with $x_0 \in (-1,1)$ and let $\delta > 0$ with $(x_0 - \delta, x_0 + \delta) \subset (-1,1)$. Then

$$\begin{aligned} K_0 &:= \{(x,y) \in K \,|\, x \in (x_0 - \delta, x_0 + \delta)\} \\ &= \{(x, \sqrt{1-x^2}) \in \mathbb{R}^2 \mid x \in (x_0 - \delta, x_0 + \delta)\} \cup \{(x, -\sqrt{1-x^2}) \in \mathbb{R}^2 \mid x \in (x_0 - \delta, x_0 + \delta)\} \\ &= \mathrm{Graph}(f_1) \cup \mathrm{Graph}(f_2), \end{aligned}$$

where $f_1, f_2 : (x_0 - \delta, x_0 + \delta) \to \mathbb{R}$, $f_1(x) = \sqrt{1-x^2}$, $f_2(x) = -\sqrt{1-x^2}$. By further restricting to points lying in the upper or lower half-plane of $\mathbb{R}^2$, we can only study one of the two arcs of K that are obtained in this way. Thus, we say that K is locally given as a graph near (x_0, y_0) or that *we can resolve* $x^2 + y^2 = 1$ *by* x *near* (x_0, y_0).. As a graph is the image of an interval of an injective continuous map, this formalizes the notion that K "looks one-dimensional".
Note that this approach fails in the points $(1,0), (-1,0) \in K$. In these points, one might however swap the roles of x and y and resolve the equation $x^2 + y^2 = 1$, i.e. write K locally as a graph of a function in y. One might now ask if this approach also works for other equations than $x^2 + y^2 = 1$, which leads to the following question:

Given a differentiable map $f : \mathbb{R}^2 \to \mathbb{R}$ *and* $c \in \mathbb{R}$, *let* $M \subset \mathbb{R}^2$ *denote the solution set of*

$$f(x,y) = c.$$

Let $(x_0, y_0) \in M$. *Can we find conditions on* f, *under which there are* $\delta, \varepsilon > 0$ *and a continuous map* $g : (x_0 - \delta, x_0 + \delta) \to (y_0 + \varepsilon, y_0 + \varepsilon)$, *such that*

$$M \cap \big((x_0 - \delta, x_0 + \delta) \times (y_0 - \varepsilon, y_0 + \varepsilon)\big) = \mathrm{Graph}(g) \qquad ?$$

Let us discuss another motivating problem for the implicit function theorem, which a priori has nothing to do with the previous example. Consider the following system of equations:

$$\begin{cases} \cos(\lambda x) + \lambda^2 e^y &= 0, \\ \sin(\lambda x) + e^y xy &= 0, \end{cases} \tag{13.8}$$

which we view as a system of equations in the variables $x, y \in \mathbb{R}$ which is parametrized by $\lambda \in \mathbb{R}$. By meditating over the equations in the case $\lambda = 1$, one observes that a solution of the system is given by $x = \pi$ and $y = 0$. Since all functions involved in (13.8) are differentiable, it is a natural question to ask whether perturbing the parameter λ by a bit away from 1, will we again obtain a solution of (13.8) that is close to $(\pi, 0)$ and if so, if the solution depends differentiably on the parater. More precisely, we ask the following question.

If we modify the parameter by taking $\lambda \in (1-\varepsilon, 1+\varepsilon)$ for some small $\varepsilon > 0$, are there continuous/differentiable functions $x, y : (1-\varepsilon, 1+\varepsilon) \to \mathbb{R}$ with $x(1) = \pi$, $y(1) = 0$ and such that $(x(\lambda), y(\lambda))$ solves (13.8) *for each $\lambda \in (1-\varepsilon, 1+\varepsilon)$?*

We shall see below that both questions can be treated as special cases of the more general implicit function theorem.

Throughout the following, we fix numbers $m, n \in \mathbb{N}$ with $m < n$.

We will further occasionally write $\mathbb{R}^n$ as $\mathbb{R}^n = \mathbb{R}^{n-m} \times \mathbb{R}^m$ and write $h \in \mathbb{R}^n$ as $h = (h_1, h_2)$, where $h_1 \in \mathbb{R}^{n-m}$, $h_2 \in \mathbb{R}^m$.

Let $U \subset \mathbb{R}^n$ open, $a \in U$ and let $f : U \to \mathbb{R}^m$ be differentiable in a. For each $h = (h_1, h_2) \in \mathbb{R}^n$ it holds by linearity of the differential that

$$Df_a(h) = Df_a(h_1, 0) + Df_a(0, h_2).$$

Thus, if we define define linear maps $D_1 f_a : \mathbb{R}^{n-m} \to \mathbb{R}^m$ and $D_2 f_a : \mathbb{R}^m \to \mathbb{R}^m$ by

$$D_1 f_a(h_1) := Df_a(h_1, 0) \qquad D_2 f_a(h_2) := Df_a(0, h_2) \quad \forall h_1 \in \mathbb{R}^{n-m},\ h_2 \in \mathbb{R}^m,$$

then

$$Df_a(h_1, h_2) = D_1 f_a(h_1) + D_2 f_a(h_2) \quad \forall (h_1, h_2) \in \mathbb{R}^{n-m} \times \mathbb{R}^m = \mathbb{R}^n.$$

We shall employ this notation throughout the rest of this section.

Before tackling the implicit function theorem, we make a general observation. If a solution set of the form $f(x, y) = c$ is given by the graph of a differentiable function, then the differential of that function in a point is fully determined by the differentials of f as the following theorem shows.

Theorem 13.8. *Let $U_1 \subset \mathbb{R}^{n-m}$ and $U_2 \subset \mathbb{R}^m$ be open. Let $f : U_1 \times U_2 \to \mathbb{R}^m$ and $g : U_1 \to U_2$ be given, so that there exists $c \in \mathbb{R}^m$ with*

$$f(x, g(x)) = c \qquad \forall x \in U_1.$$

If g is differentiable in $a \in U_1$, f is differentiable in $(a, g(a))$ and $D_2 f_{(a,g(a))} : \mathbb{R}^m \to \mathbb{R}^m$ is an isomorphism, then

$$Dg_a = -(D_2 f_{(a,g(a))})^{-1} \circ D_1 f_{(a,g(a))}.$$

Proof. We observe that $f(x, g(x)) = (f \circ (\mathrm{id}_{\mathbb{R}^m}, g))(x)$ for all $x \in U_1$, which is a composition of differentiable maps, hence itself differentiable in a. From the identity $f(x, g(x)) = c$ we thus obtain for all $h \in \mathbb{R}^{n-m}$ by the chain rule and Theorem 12.22 that

$$\begin{aligned} 0 &= D(f \circ (\mathrm{id}_{\mathbb{R}^m}, g))_a(h = (Df_{(a,g(a))} \circ (\mathrm{id}_{\mathbb{R}^m}, Dg_a))(h) \\ &= Df_{(a,g(a))}(h, Dg_a(h)) = D_1 f_{(a,g(a))}(h) + D_2 f_{(a,g(a))}(Dg_a(h)). \end{aligned}$$

This implies $D_2 f_{(a,g(a))}(Dg_a(h)) = -D_1 f_{(a,g(a))}(h)$ and since $D_2 f_{(a,g(a))}$ is invertible, it follows that

$$Dg_a(h) = -(D_2 f_{(a,g(a))})^{-1}(D_1 f_{(a,g(a))}(h)) \qquad \forall h \in \mathbb{R}^{n-m}.$$

□

The previous theorem assumes the existence of a differentiable map g with $f(x, g(x)) = c$, b ut it is of course not at all clear at this point which such a g exists. Such g is also called an *implicit function*, because it is not explicitly defined by a mathematical expression, but rather *implicitly* by demanding it to satisfy $f(x, g(x)) = c$. The implicit function theorem now gives some handy criteria for the existence of such a map g.

Theorem 13.9 (Implicit function theorem)**.** *Let $U \subset \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}^m$ be continuously differentiable and let $(a, b) \in U$, where $a \in \mathbb{R}^{n-m}$ and $b \in \mathbb{R}^m$. Put $c := f(a, b)$. If $D_2 f_{(a,b)} : \mathbb{R}^m \to \mathbb{R}^m$ is an isomorphism, then there are open neighborhoods $U_1 \subset \mathbb{R}^{n-m}$ of a and $U_2 \subset \mathbb{R}^m$ of b with $U_1 \times U_2 \subset U$ and a continuously differentiable map $g : U_1 \to U_2$ with $g(a) = b$ so that*

$$f(x, g(x)) = c \qquad \forall x \in U_1$$

and so that if $(x, y) \in U_1 \times U_2$ satisfies $f(x, y) = c$, then $y = g(x)$. Moreover, the differential of g in a is given by

$$Dg_a = -(D_2 f_{(a,b)})^{-1} \circ D_1 f_{(a,b)}.$$

Proof. We assume w.l.o.g. that $c = 0$, since in the general case, we might replace f by $f_1(x, y) = f(x, y) - c$ and carry out the line of argument for f_1 instead. The proof is an application of the inverse function theorem. Let

$$F : U \to \mathbb{R}^{n-m} \times \mathbb{R}^m = \mathbb{R}^n, \qquad F(x, y) = (x, f(x, y)).$$

Then F is continuously differentiable and we compute for all $(x, y) \in U$, $h_1 \in \mathbb{R}^{n-m}$ and $h_2 \in \mathbb{R}^m$ that

$$DF_{(x,y)}(h_1, h_2) = (h_1, D_1 f_{(x,y)}(h_1) + D_2 f_{(x,y)}(h_2)).$$

We claim that $DF_{(a,b)}$ is an isomorphism. Let $(k_1, k_2) \in \mathbb{R}^{n-m} \times \mathbb{R}^m$ be arbitrary. Since $D_2 f_{(a,b)}$ is an isomorphism, there exists $k_3 \in \mathbb{R}^m$ with $D_2 f_{(a,b)}(k_3) = k_2 - D_1 f_{(a,b)}(k_1) - D_2 f_{(x,y)}(k_2)$. Then

$$\begin{aligned} DF_{(a,b)}(k_1, k_2 + k_3) &= DF_{(a,b)}(k_1, k_2) + DF_{(a,b)}(0, k_3) \\ &= (k_1, D_1 f_{(a,b)}(k_1) + D_2 f_{(a,b)}(k_2)) + (0, k_2 - D_1 f_{(a,b)}(k_1) - D_2 f_{(a,b)}(k_2)) \\ &= (k_1, k_2), \end{aligned}$$

hence $(k_1, k_2) \in \operatorname{im} DF_{(a,b)}$. This shows that $DF_{(a,b)}$ is surjective, hence an isomorphism by Corollary 7.23.c). Thus, by the inverse function theorem, there are open neighborhoods $U_0 \subset \mathbb{R}^n$

of (a,b) and $V_0 \subset \mathbb{R}^n$ of $F(a,b) = (a,c)$, such that $F|_{U_0} : U_0 \to V_0$ is bijective and its inverse $F^{-1} : V_0 \to U_0$ is continuously differentiable. By definition of F, the inverse is of the form

$$F^{-1}(u,v) = (u, h(u,v)),$$

for some suitable continuously differentiable $h : V_0 \to \mathbb{R}^m$. By construction, for $(x,y) \in U_0$ we obtain

$$f(x,y) = 0 \quad \Leftrightarrow \quad F(x,y) = (x,0) \quad \Leftrightarrow \quad y = h(x,0). \tag{13.9}$$

In particular, $h(a,0) = b$. Since h is continuous, there are open neighborhooda $U_1 \subset \mathbb{R}^{n-m}$ of a and $U_2 \subset \mathbb{R}^m$ of b with $U_1 \times U_2 \subset U_0$, such that $h(x,0) \in U_2$ for all $x \in U_1$. Define $g : U_1 \to U_2$, $g(x) = h(x,0)$. Then g is continuously differentiable with $g(a) = b$ and by (13.9) it has the desired property. The formula for Dg_a then follows from Theorem 13.8. □

Remark 13.10. In the situation of Theorem 13.9, the condition that $D_2 f_{(a,b)}$ is an isomorphism can be checked from the Jacobi matrix of f in (a,b). By definition, $Df_{(a,b)}(v) = J_f(a,b) \cdot v$ for all $v \in \mathbb{R}^n$ and one observes from the definition of $D_2 f_{(a,b)}$ that $D_2 f_{(a,b)}(v) = J_{2,f}(a,b) \cdot v$ for all $v \in \mathbb{R}^m$, where, with $f = (f_1, \dots, f_m)$,

$$J_{2,f}(a,b) \in M(m \times m, \mathbb{R}), \quad J_{2,f}(a,b) = \begin{pmatrix} \partial_{n-m+1} f_1(a,b) & \partial_{n-m+2} f_1(a,b) & \dots & \partial_n f_1(a,b) \\ \partial_{n-m+1} f_2(a,b) & \partial_{n-m+2} f_2(a,b) & \dots & \partial_n f_2(a,b) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{n-m+1} f_m(a,b) & \partial_{n-m+2} f_m(a,b) & \dots & \partial_n f_m(a,b) \end{pmatrix},$$

i.e. the columns of $J_{2,f}(a,b)$ are the last m columns of $J_f(a,b)$. So the condition for the implicit function theorem to be applicable is that $J_{2,f}(a,b)$ is invertible, which is in turn equivalent to

$$\det J_{2,f}(a,b) \neq 0.$$

Since the case of $m = 1$ is very common in applications of the implicit function theorem, we state it explicitly.

Corollary 13.11. *Let $U \subset \mathbb{R}^n$ be an open subset and let $f : U \to \mathbb{R}$ be continuously differentiable. Let $a = (a_1, \dots, a_n) \in U$ with $f(a) = 0$. If $\partial_n f(a) \neq 0$, then there is an open neighborhood U' of $a' := (a_1, \dots, a_{n-1}) \in \mathbb{R}^{n-1}$ and a continuously differentiable map $g : U' \to \mathbb{R}$, such that*

$$f(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})) = 0 \qquad \forall (x_1, \dots, x_{n-1}) \in U'.$$

Moreover,

$$\nabla g(a') = \frac{1}{\partial_n f(a)} (\partial_1 f(a), \partial_2 f(a), \dots, \partial_{n-1} f(a)).$$

Example 13.12. We consider again the solution set K of

$$x^2 + y^2 = 1.$$

For $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2 + y^2 - 1$, it holds that $K = \{(x,y) \in \mathbb{R}^2 \mid f(x,y) = 0\}$. As a polynomial, f is continuously differentiable with $\partial_1 f(x,y) = 2x$ and $\partial_2 f(x,y) = 2y$ for all $(x,y) \in \mathbb{R}^2$. Hence, if $(x_0, y_0) \in K$ with $y_0 \neq 0$, it holds that $\partial_2 f(x_0, y_0) \neq 0$. Thus by the implicit function theorem there are $\varepsilon, \delta > 0$ and a continuously differentiable map $g : (x_0 - \delta, x_0 + \delta) \to (y_0 - \varepsilon, y_0 + \varepsilon)$ such that $K \cap ((x_0 - \delta, x_0 + \delta) \times (y_0 - \varepsilon, y_0 + \varepsilon)) = \{(x, g(x)) \in \mathbb{R}^2 \mid x \in (x_0 - \delta, x_0 + \delta)\} = \mathrm{Graph}(g)$. As we have seen above, in this particular example the map g is either given by $f_1(x) = \sqrt{1 - x^2}$ or by $f_2(x) = -\sqrt{1 - x^2}$, depending on the sign of y_0.

Example 13.13. We reconsider the system of equations (13.8) and want to apply the implicit function theorem to answer the above question.
Let $f : \mathbb{R}^3 \to \mathbb{R}^2$, $f(\lambda, x, y) = (\cos(\lambda x) + \lambda^2 e^y, \sin(\lambda x) + e^y xy)$. Then (λ, x, y) is a solution of (13.8) if and only if $f(\lambda, x, y) = (0, 0)$. f is continuously differentiable and we compute its Jacobi matrix at an arbitrary point with $f = (f_1, f_2)$ as

$$\begin{aligned} J_f(\lambda, x, y) &= \begin{pmatrix} \partial_1 f_1(\lambda, x, y) & \partial_2 f_1(\lambda, x, y) & \partial_3 f_1(\lambda, x, y) \\ \partial_1 f_2(\lambda, x, y) & \partial_2 f_2(\lambda, x, y) & \partial_3 f_2(\lambda, x, y) \end{pmatrix} \\ &= \begin{pmatrix} -x\sin(\lambda x) & -\lambda\sin(\lambda x) & \lambda^2 e^y \\ x\cos(\lambda x) & \lambda\cos(\lambda x) + e^y y & xe^y(1+y) \end{pmatrix}. \end{aligned}$$

For $(h_1, h_2, h_3) \in \mathbb{R}^3$ we let $D_1 f_{(\lambda,x,y)} : \mathbb{R} \to \mathbb{R}^2$ and $D_2 f_{(\lambda,x,y)} : \mathbb{R}^2 \to \mathbb{R}^2$ be given so that

$$Df_{(\lambda,x,y)}(h_1, h_2, h_3) = D_1 f_{(\lambda,x,y)}(h_1) + D_2 f_{(\lambda,x,y)}(h_2, h_3).$$

In the notation of Remark 13.10 it holds that

$$D_2 f_{(\lambda,x,y)}(h_2, h_3) = J_{2,f}(\lambda, x, y) \cdot \begin{pmatrix} h_2 \\ h_3 \end{pmatrix} \qquad \forall (h_2, h_3) \in \mathbb{R}^2,$$

where $J_{2,f}(\lambda, x, y) = \begin{pmatrix} -\lambda\sin(\lambda x) & \lambda^2 e^y \\ \lambda\cos(\lambda x) + e^y y & xe^y(1+y) \end{pmatrix} \in M(2 \times 2, \mathbb{R})$. We have observed above that $f(1, \pi, 0) = (0, 0)$. We compute that

$$J_{2,f}(1, \pi, 0) = \begin{pmatrix} 0 & 1 \\ -1 & \pi \end{pmatrix},$$

hence $\det J_{2,f}(1, \pi, 0) = 1 \neq 0$, so $D_2 f_{(1,\pi,0)}$ is an isomorphism. The implicit function theorem thus implies:

There are $\varepsilon > 0$, an open neighborhood $U \subset \mathbb{R}^2$ of $(\pi, 0)$ and a continuously differentiable map $g : (1-\varepsilon, 1+\varepsilon) \to U$, such that $g(1) = (\pi, 0)$,

$$f(\lambda, g(\lambda)) = (0, 0) \qquad \forall \lambda \in (1-\varepsilon, 1+\varepsilon),$$

and such that every $(\lambda, x, y) \in (1-\varepsilon, 1+\varepsilon) \times U$ with $f(\lambda, x, y) = (0, 0)$ satisfies $(x, y) = g(\lambda)$. Let $(x(\lambda), y(\lambda)) := g(\lambda)$. Then we have shown in particular that there is an $\varepsilon > 0$, such that for each $\lambda \in (1-\varepsilon, 1+\varepsilon)$ the system (13.8) has a unique solution $(x(\lambda), y(\lambda)) \in U$ (but might have more solutions outside of U).
Similar as in the previous section, we can compute approximate solutions using the implicit function theorem as well. By the formula from the implicit function theorem, it holds that

$$g'(1) = -(D_2 f_{(1,\pi,0)})^{-1}(D_1 f_{(1,\pi,0)}(1)) = -\begin{pmatrix} 0 & 1 \\ -1 & \pi \end{pmatrix}^{-1} \begin{pmatrix} -2 \\ \pi \end{pmatrix} = -\begin{pmatrix} \pi & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -2 \\ \pi \end{pmatrix} = \begin{pmatrix} 3\pi \\ 2 \end{pmatrix},$$

where we used Corollary 9.27. By the Péano form of Taylor's theorem applied to g in 1, it holds that

$$g(1+\mu) = g(1) + Dg_1(\mu) + \varphi(\mu) = g(1) + g'(1)\mu + \varphi(\mu) = \begin{pmatrix} \pi + 3\pi\mu \\ 2\mu \end{pmatrix} + \varphi(\mu),$$

where $\varphi : (-\varepsilon, \varepsilon) \to \mathbb{R}^2$ satisfies $\lim_{\mu \to 0} \frac{\|\varphi(\mu)\|}{\|\mu\|} = 0$. With $\varphi = (\varphi_1, \varphi_2)$, this reads as

$$\begin{cases} x(1+\mu) &= \pi + 3\pi\mu + \varphi_1(\mu), \\ y(1+\mu) &= 2\mu + \varphi_2(\mu), \end{cases}$$

for the approximate solutions of the system.

13.4 The Lagrange multiplier theorem

This section is not discussed in the lecture videos.

Assume that we consider the universe at a fixed moment in time. Which point on the surface of the earth is closest to the center of the sun? If we imagine the surface of the earth as

$$S = \{(x,y,z) \in \mathbb{R}^3 \mid \|(x,y,z)\| = 1\} = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

and the center of the sun as a point $P = (a,b,c) \in \mathbb{R}^3$ with $\|P\| > 1$, then consider the function

$$f : \mathbb{R}^3 \to \mathbb{R}, f(x,y,z) = \|(x,y,z) - (a,b,c)\|^2 = (x-a)^2 + (y-b)^2 + (z-c)^2.$$

The point we are searching for will not be a minimum point of f, but a minimum point of $f|_S$. While f is differentiable as However, we can not apply our methods for finding local extrema since S is not open in $\mathbb{R}^3$. S is easily seen to be closed and bounded and thus compact by the Heine-Borel theorem, so $f|_S$ attains its minimum and its maximum. But how do we find it? The motivating question for this section is thus:

Which point in $S = \{p \in \mathbb{R}^3 \mid \|p\| = 1\}$ has the shortest distance to a given $P = (a,b,c) \in \mathbb{R}^3 \setminus S$?

We want to obtain a method for finding *conditional local extrema* of a function on an open subset of $\mathbb{R}^n$, i.e. we want to find the minimum value of f under the condition that we only consider points that satisfy a given equation. For this purpose, the following result can be derived from the implicit function theorem.

Theorem 13.14 (Lagrange multiplier theorem)**.** *Let $m,n \in \mathbb{N}$ with $m < n$ and let $U \subset \mathbb{R}^n$ be open. Let $f, \varphi_1, \dots, \varphi_m : U \to \mathbb{R}$ be continuously differentiable and let $\varphi := (\varphi_1, \dots, \varphi_m) : U \to \mathbb{R}^m$. Assume that* $\operatorname{rank} D\varphi_x = m$ *for all $x \in U$ and let $M := \{x \in U \mid \varphi(x) = 0\}$. If $f|_M$ attains its maximum or its minimum in $x_0 \in M$, then there exist $\lambda_1, \dots, \lambda_m \in \mathbb{R}$, such that*

$$\nabla f(x_0) = \sum_{i=1}^{m} \lambda_i \nabla \varphi_i(x_0).$$

Proof. Let $a \in M$ be a local maximum point or a local minimum point of $f|_M$. By assumption

$$m = \operatorname{rank} D\varphi_x = \operatorname{rank} J_\varphi(x) = \operatorname{rank} \begin{pmatrix} \partial_1\varphi_1(a) & \partial_2\varphi_1(a) & \dots & \partial_n\varphi_1(a) \\ \partial_1\varphi_2(a) & \partial_2\varphi_2(a) & \dots & \partial_n\varphi_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1\varphi_m(a) & \partial_2\varphi_m(a) & \dots & \partial_n\varphi_m(a) \end{pmatrix}.$$

Thus, by definition of the rank, $J_\varphi(x)$ has m linearly independent columns and we can assume w.l.o.g. (since this can be achieved by simply relabeling the coordinates) that the last m columns of the matrix are linearly independent, i.e. that

$$J_{2,\varphi}(a) = \begin{pmatrix} \partial_{n-m+1}\varphi_1(a) & \dots & \partial_n\varphi_1(a) \\ \partial_{n-m+1}\varphi_2(a) & \dots & \partial_n\varphi_2(a) \\ \vdots & \ddots & \vdots \\ \partial_{n-m+1}\varphi_m(a) & \dots & \partial_n\varphi_m(a) \end{pmatrix} \in M(m \times m, \mathbb{R})$$

is invertible. By Remark 13.10, this shows that with $D_2\varphi_a : \mathbb{R}^m \to \mathbb{R}^m$ defined in analogy with the previous section, this shows that $D_2\varphi_a$ is an isomorphism. Let $a = (a_1, a_2)$, where $a_1 \in \mathbb{R}^{n-m}$

and $a_2 \in \mathbb{R}^m$. Then by the implicit function theorem, there are open neighborhoods $U_1 \subset \mathbb{R}^{n-m}$ of a_1 and $U_2 \subset \mathbb{R}^m$ of a_2 and a differentiable map $g : U_1 \to U_2$ with $g(a_1) = a_2$, such that

$$M \cap (U_1 \times U_2) = \{(x, y) \in U_1 \times U_2 \mid \varphi(x, y) = 0\} = \text{Graph}(g).$$

Let

$$F : U_1 \to \mathbb{R}, \qquad F(x) = f(x, g(x)).$$

Assume w.l.o.g. that a is a local *maximum* point of $f|_M$. We further assume that U_1 and U_2 are chosen such that $f(x, y) \leq f(a)$ for all $(x, y) \in M \cap (U_1 \cap U_2)$. Then it particularly follows that $F(x) \leq F(a_1)$ for all $x \in U_1$.
Hence, a_1 is a local maximum point of F (without any constraints!). So by Theorem 12.36, it holds for all $h \in \mathbb{R}^{n-m}$ that

$$0 = DF_{a_1}(h) = D(f \circ (\text{id}, g))_{a_1}(h) = Df_a(h, Dg_{a_1}(h)) = D_1 f_a(h) + D_2 f_a(Dg_{a_1}(h)),$$

where we used the chain rule. By the implicit function theorem and the definition of g, it further holds that

$$Dg_{a_1} = -(D_2\varphi_a)^{-1} \circ D_1\varphi_a.$$

Inserting this into the above equation, we obtain that

$$D_1 f_a = D_2 f_a \circ (D_2\varphi_a)^{-1} \circ D_1\varphi_a = \Lambda \circ D_1\varphi_a,$$

where $\Lambda \in L(\mathbb{R}^m, \mathbb{R})$ is given by $\Lambda := D_2 f_a \circ (D_2\varphi_a)^{-1}$. Moreover,

$$D_2 f_a = D_2 f_a \circ (D_2\varphi_a)^{-1} \circ D_2\varphi_a = \Lambda \circ D_2\varphi_a.$$

Combining the last equation, we obtain for all $h_1 \in \mathbb{R}^{n-m}$ and $h_2 \in \mathbb{R}^m$ that

$$\begin{aligned} Df_a(h_1, h_2) &= D_1 f_a(h_1) + D_2 f_a(h_2) = \Lambda(D_1\varphi_a(h_1)) + \Lambda(D_2\varphi_a(h_2)) \\ &= \Lambda(D_1\varphi_a(h_1)) + D_2\varphi_a(h_2)) = \Lambda(D\varphi_a(h_1, h_2)), \end{aligned}$$

hence $Df_a = \Lambda \circ D\varphi_a$. By the Riesz representation theorem, there exists $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ with $\Lambda(v_1, \ldots, v_n) = \sum_{i=1}^n \lambda_i v_i$ for all $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$. Since $D\varphi_a = (D\varphi_1)_a, \ldots, (D\varphi_m)_a)$ by Theorem 12.14, this shows that

$$Df_a(v) = \sum_{i=1}^n \lambda_i (D\varphi_i)_a.$$

Using the connection between differentials and gradients of functions, one directly derives the claim. □

Remark 13.15. (1) The numbers $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ in the situation of Theorem 13.14 are called the *Lagrange multipliers* of the problem.

(2) The Lagrange multiplier theorem has important applications to classical mechanics, more precisely in the Lagrangian formalism of classical mechanics. Assume an imaginary physical object is moving in $\mathbb{R}^n$, but underlying some constraint, like a small ball in $\mathbb{R}^3$ that is rolling in a bowl with the constraint that the ball is attached to the bowl. Such constraints can usually be described as conditions of the form $\varphi(x) = 0$, where φ is continuously differentiable. Associated with such constrainst are constraint forces acting upon that object. The gradients of the φ_i point in the direction of the constraint force associated with the constraint that $\varphi_i(x) = 0$. The Lagrange multipliers then roughly describe the strengths of the constraint forces and how the total constraint force acting upon an object is "composed" of the constraint forces of the individual constraints.

(3) In nonlinear optimization, there is a generalization of the Lagrange multiplier theorem, called the Karush-Kuhn-Tucker theorem (or KKT theorem), which helps in finding conditional local extrema under conditions of the form $\varphi_i(x) \leq 0$, i.e. handling inequalities as well. While this is a natural situation in applications e.g. to economics, the proof and formulation of the result are more intricate, so we shall not discuss it here.

Example 13.16. As in the beginning of this section, we consider $P = (a, b, c) \in \mathbb{R}^3$ with $\|P\| > 1$ and

$$f : \mathbb{R}^3 \to \mathbb{R}, \qquad f(x, y, z) = (x - a)^2 + (y - b)^2 + (z - c)^2.$$

We want to find the minimum of f on S, where $S = \{(x, y, z) \mid \varphi(x, y, z) = 0\}$ with

$$\varphi : \mathbb{R}^3 \to \mathbb{R}, \qquad \varphi(x, y, z) = x^2 + y^2 + z^2 - 1.$$

Both f and φ are polynomials, hence continuously differntiable. Moreover, $\nabla\varphi(x, y, z) = (2x, 2y, 2z)$, so $\nabla\varphi(x, y, z) \neq 0$ if $(x, y, z) \neq 0$. In particular, this shows that $D\varphi_{(x,y,z)} : \mathbb{R}^3 \to \mathbb{R}$ satisfies $D\varphi_{(x,y,z)} \neq 0$ for all $(x, y, z) \in S$, so for dimensional reasons $\operatorname{rank} D\varphi_{(x,y,z)} = 1$ for all $(x, y, z) \in S$.
Hence, by the Lagrange multiplier theorem, for every local minimum point of $f|_S$ there exists $\lambda \in \mathbb{R}$ with

$$\nabla f(x, y, z) = \lambda \nabla g(x, y, z).$$

Since $\nabla f(x, y, z) = (2(x - a), 2(y - b), 2(z - c))$ for all $(x, y, z) \in \mathbb{R}^3$, this amounts to

$$\begin{cases} 2(x - a) &= 2\lambda x, \\ 2(y - b) &= 2\lambda y, \\ 2(z - c) &= 2\lambda z, \end{cases} \qquad \Leftrightarrow \qquad \begin{cases} (1 - \lambda)x &= a, \\ (1 - \lambda)y &= b, \\ (1 - \lambda)z &= c, \end{cases}$$

Since $P = (a, b, c) \neq 0$, the system has no solution for $\lambda = 1$, so assume that $\lambda \neq 1$. Then we need to find λ with

$$x = \frac{a}{1 - \lambda}, \qquad y = \frac{b}{1 - \lambda}, \qquad z = \frac{c}{1 - \lambda} \tag{13.10}$$

and such that $(x, y, z) \in S$. Thus, we need to find λ with

$$1 = x^2 + y^2 + z^2 = \left(\frac{a}{1 - \lambda}\right)^2 + \left(\frac{b}{1 - \lambda}\right)^2 + \left(\frac{c}{1 - \lambda}\right)^2 = \frac{a^2 + b^2 + c^2}{(1 - \lambda)^2} = \frac{\|P\|^2}{(1 - \lambda)^2}.$$

Hence, $(1 - \lambda)^2 = \|P\|^2$, which has the two solutions $\lambda_1 = 1 - \|P\|$ and $\lambda_2 = 1 + \|P\|$. Inserting this into (13.10), we obtain the two possible local extremum points

$$p_1 = \frac{1}{\|P\|}(a, b, c), \qquad p_2 = -\frac{1}{\|P\|}(a, b, c).$$

Since $f|_S$ attains its maximum and its minimum, one of these points must be a local maximum point and the other be a local minimum point. We compute that

$$\begin{aligned} f(p_1) &= (\frac{1}{\|P\|})a - a)^2 + (\frac{1}{\|P\|})b - b)^2 + (\frac{1}{\|P\|})c - c)^2 \\ &= \left(1 - \frac{1}{\|P\|}\right)^2 a^2 + \left(1 - \frac{1}{\|P\|}\right)^2 b^2 + \left(1 - \frac{1}{\|P\|}\right)^2 c^2 \\ &= \frac{(\|P\| - 1)^2}{\|P\|^2}(a^2 + b^2 + c^2) = (\|P\| - 1)^2, \end{aligned}$$

and analogously $f(p_2) = (\|P\| + 1)^2$. Since $\|P\| > 1$, this shows that the minimum of $f|_S$ is given by $f(p_1) = (\|P\| - 1)^2$, so the minimum distance from P to S is attained in $\frac{1}{\|P\|}(a, b, c)$ and its value is $\sqrt{f(p_1)} = \|P\| - 1$.

Example 13.17. Consider

$$M = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1, x = z\}$$

and let $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x + y$. We want to show that $f|_M$ attains its maximum and compute it.
For all $(x, y, z) \in M$ we compute from the definition of M that

$$\|(x, y, z)\|^2 = x^2 + y^2 + z^2 = 2x^2 + y^2 \leq 2(x^2 + y^2) \leq 2.$$

Hence, $M \subset \overline{B}_{\sqrt{2}}(0)$, so M is bounded. Moreover, one easily checks that M is closed, hence M is compact by the Heine-Borel theorem. As a polynomial, f is continuously differentiable, hence $f|_M$ is continuous and attains its maximum and its minimum. We write

$$M = \{(x, y, z) \in \mathbb{R}^2 \mid \varphi(x, y, z) = (0, 0)\},$$

where

$$\varphi : \mathbb{R}^3 \to \mathbb{R}^2, \qquad \varphi(x, y, z) = (x^2 + y^2 - 1, x - z).$$

The components of φ are polynomials, hence φ is continuously differentiable. We compute that

$$J_\varphi(x, y, z) = \begin{pmatrix} 2x & 2y & 0 \\ 1 & 0 & -1 \end{pmatrix} \qquad \forall (x, y, z) \in \mathbb{R}^3,$$

from which one easily checks that $\operatorname{rank} J_\varphi(x, y, z) = 2$ whenever $(x, y, z) \neq (0, 0, 0)$. In particular,

$$\operatorname{rank} D\varphi_{(x,y,z)} = \operatorname{rank} J_\varphi(x, y, z) = 2 \quad \forall (x, y, z) \in M,$$

as obviously $(0, 0, 0) \notin M$. Thus, we can apply the Lagrange multiplier theorem, which says that for every local maximum point or local minimum point of $f|_M$ there are $\lambda_1, \lambda_2 \in \mathbb{R}$ with

$$\begin{aligned} & \nabla f(x, y, z) = \lambda_1 \nabla \varphi_1(x, y, z) + \lambda_2 \nabla \varphi_2(x, y, z) \\ \Leftrightarrow \quad & \partial_i f(x, y, z) = \lambda_1 \partial_i \varphi_1(x, y, z) + \lambda_2 \partial_i \varphi_2(x, y, z) \quad \forall i \in \{1, 2, 3\}, \end{aligned} \tag{13.11}$$

where $\varphi = (\varphi_1, \varphi_2)$. The partial derivatives of φ_1 and φ_2 are read off from $J_\varphi(x, y, z)$, while $\nabla f(x, y, z) = (1, 1, 0)$ for all $(x, y, z). \in \mathbb{R}^3$. Hence, (13.11) holds if and only if

$$\begin{cases} 1 &= 2\lambda_1 x + \lambda_2, \\ 1 &= 2\lambda_2 y \\ 0 &= \lambda_2 \end{cases}$$

It follows immediately that $\lambda_1 \neq 0$, $\lambda_2 = 0$ and $x = y = \frac{1}{2\lambda_1}$. Since $(x, y, z) \in M$, it further needs to hold that

$$1 = x^2 + y^2 = \frac{1}{4\lambda_1^2} + \frac{1}{4\lambda_1^2} \quad \Leftrightarrow \quad \lambda_1^2 = \frac{1}{2} \quad \Leftrightarrow \quad \lambda_1 = \frac{1}{\sqrt{2}} \ \vee \ \lambda_1 = -\frac{1}{\sqrt{2}}.$$

Together with $x = z$, this implies that the only possible local minimum or local maximum points of $f|_M$ are

$$p_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad p_2 = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right).$$

Since $f(p_1) = 2 \cdot \frac{1}{\sqrt{2}} = \sqrt{2}$ and $f(p_2) = -\sqrt{2}$, it follows that the maximum of $f|_M$ is $\sqrt{2}$ and that it is attained in $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$.

This concludes our collection of important theorems from higher-dimensional calculus and thereby the lecture course. Another important application of the Banach fixed point theorem is the theorem of Picard-Lindelöf about the existence and uniqueness of solutions of ordinary differential equations. This result will very likely be discussed in Mathematics 3 in the upcoming semester.